Econometrics Spring School 2016
Econometric Modelling

Jurgen A Doornik, David F. Hendry, and Felix Pretis

George-Washington University

March 2016

Lecture 6:
Model selection theory and evidence
Introduction to Monte Carlo Simulation

Economies high dimensional, interdependent, heterogeneous, and evolving: comprehensive specification of all events is impossible.

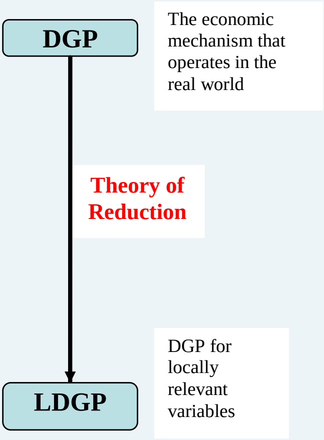
Data generation process (DGP): joint density of all variables in economy.
Impossible to accurately theorize about or model precisely:
Too high dimensional and far too non-stationary.

Need to reduce to manageable size in 'local DGP' (LDGP):
the DGP in space of variables under analysis.

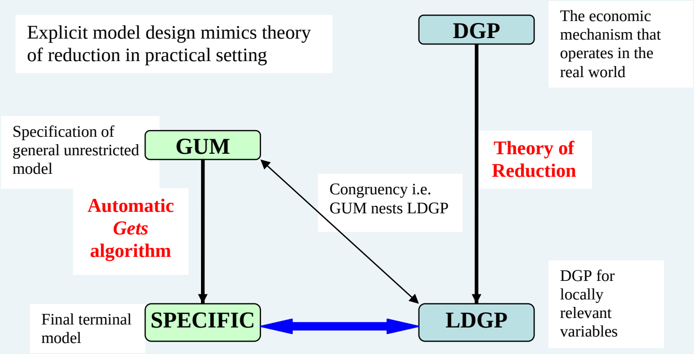
Models reflect LDGP, not copies: designed to satisfy selection criteria.
Knowing LDGP, can generate 'look alike data' which only deviate from actual data by unpredictable noise.

Therefore LDGP is the target for model selection.

DGP to LDGP


DGP
The economic mechanism that operates in the real world
Theory of Reduction
LDGP
DGP for locally relevant variables

GUM to selected model



Simulations in OxMetrics

- PCNaive (OxMetrics Module)
- Programming using Ox

$\rightarrow$ focus on PCNaive

PCNaive

- Simple to use, intuitive menu based
- Quick Monte Carlo Simulations
- Teaching tool

Structure on 3 levels:

- AR(1) experiment (basic)
- Static experiment (intermediate)
- Advanced experiment

Using PC Naive

1. AR(1) Experiment for introduction
2. Assessing performance of automatic model selection

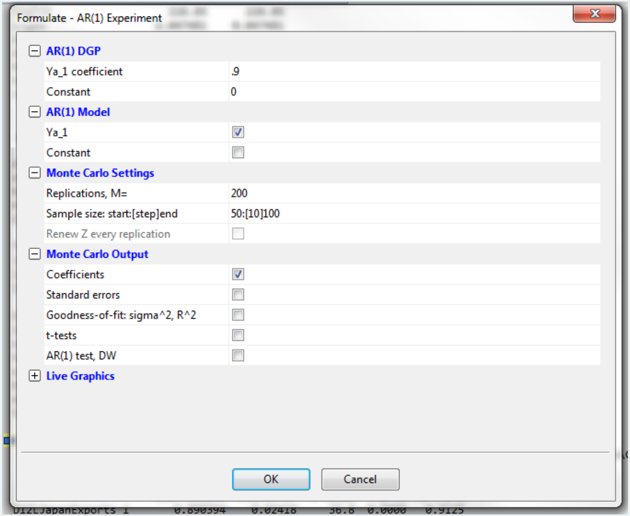
Model - Monte Carlo - AR(1) Experiment

Setup:

- AR(1) DGP, choose coefficient
- Monte Carlo Replications
- Live Graphics – Generated Data, Histograms

Institute for New Economic Thinking at the Oxford Martin School

AR(1) Experiment Overview:



Start with very simple example: DataSet1.in7

Dataset consists of 20 explanatory variables:

$$\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{20})' \sim \mathsf{IN}_{20}\left[\boldsymbol{0}, \boldsymbol{I}\right]$$

and 1 dependent variable:

$$\begin{aligned} y_t &= \beta_1 z_{1,t} + \beta_2 z_{2,t} + \beta_3 z_{3,t} + \beta_4 z_{4,t} + \beta_5 z_{5,t} + \epsilon_t \\ \epsilon_t &\sim \mathsf{IN}\left[0, 1\right] \end{aligned}$$

where $T = 100$, $\beta_1 = 0.2$, $\beta_2 = 0.3$, $\beta_3 = 0.4$, $\beta_4 = 0.5$, and $\beta_5 = 0.6$ which gives $\mathsf{E}\left(\mathsf{t}_{\beta_1}\right) = \psi_1 = 2$, $\psi_2 = 3$, $\psi_3 = 4$, $\psi_4 = 5$, $\psi_5 = 6$.

Instead of all using the same dataset, let's generate our own.

Open Modeling option, → select Monte Carlo, → select Advanced experiment.
Specify DGP as above.
We won't analyse the model so specify any model.
Important Advanced Monte Carlo settings → all use a different number of replications $M_i + 2T$.
The last draw is stored, so this will guarantee that you all have different datasets.
Save results

The general unrestricted model is:

$$y_t = \beta_0 + \sum_{i=1}^{N} \beta_i z_{i,t} + u_t$$

$$\begin{aligned}
y_t &= \underset{(0.116)}{0.189} + \underset{(0.117)}{0.225} z_{1,t} + \underset{(0.115)}{0.411} z_{2,t} + \underset{(0.119)}{0.672} z_{3,t} + \underset{(0.109)}{0.311} z_{4,t} + \\
&\quad \underset{(0.111)}{0.683} z_{5,t} + \underset{(0.12)}{0.273} z_{6,t} + \underset{(0.118)}{0.341} z_{7,t} + \underset{(0.116)}{0.081} z_{8,t} + \underset{(0.125)}{0.070} z_{9,t} + \\
&\quad \underset{(0.12)}{0.334} z_{10,t} + \underset{(0.114)}{0.020} z_{11,t} + \underset{(0.107)}{0.032} z_{12,t} + \underset{(0.112)}{0.109} z_{13,t} + \\
&\quad \underset{(0.116)}{0.113} z_{14,t} + \underset{(0.106)}{0.034} z_{15,t} + \underset{(0.117)}{0.022} z_{16,t} + \underset{(0.113)}{0.187} z_{17,t} + \\
&\quad \underset{(0.103)}{0.165} z_{18,t} + \underset{(0.101)}{0.132} z_{19,t} - \underset{(0.127)}{0.208} z_{20,t} \\
\sigma &= 1.018; \quad \mathsf{R}^2 = 0.592; \quad L = -131.84.
\end{aligned}$$

How do we select which variables matter and which do not matter?

In this example, variables are orthogonal, therefore 1-cut sufficient.

Consider a perfectly orthogonal regression model:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \qquad (1)$$

$\mathsf{E}[z_{i,t} z_{j,t}] = \lambda_{i,i}$ 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:

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

Larger values retained: all others eliminated.
Only one decision needed regardless of size of N

At $\alpha = 0.05$, $c_\alpha = 1.98$.
In our example we would retain $z_2, z_3, z_4, z_5, z_6, z_7, z_{10}$.

Errors:

- Failed to retain z_1 (a relevant variable – enters the LDGP);
- Mistakenly retained z_6, z_7, z_{10} (irrelevant variables – do not enter the LDGP).

At $\alpha = 0.01$, $c_\alpha = 2.62$.
In our example we would retain $z_2, z_3, z_4, z_5, z_7, z_{10}$.

How did we do in our Monte Carlo experiment?

First analyse retention of irrelevant variables, then consider retaining relevant variables.

Probabilities of null rejections in t-testing for N irrelevant regressors at significance level α (critical value c_α):

event	probability	retain
$\mathsf{P}\left(\lvert \mathsf{t}_i \rvert < c_\alpha,\ \forall i = 1, \ldots N\right)$	$(1-\alpha)^N$	0
$\mathsf{P}\left(\lvert \mathsf{t}_i \rvert \geq c_\alpha \mid \lvert \mathsf{t}_j \rvert < c_\alpha,\ \forall j \neq i\right)$	$N\alpha\,(1-\alpha)^{N-1}$	1
$\vdots$	$\vdots$	$\vdots$
$\mathsf{P}\left(\lvert \mathsf{t}_i \rvert < c_\alpha \mid \lvert \mathsf{t}_j \rvert \geq c_\alpha,\ \forall j \neq i\right)$	$N\alpha^{(N-1)}\,(1-\alpha)$	$N-1$
$\mathsf{P}\left(\lvert \mathsf{t}_i \rvert \geq c_\alpha,\ \forall i = 1, \ldots N\right)$	α^N	N

Average number of null variables retained is:

$$k = \sum_{i=0}^{N} i \frac{N!}{i!\,(N-i)!} \alpha^i \,(1-\alpha)^{N-i} = N\alpha. \qquad (2)$$

For $N = 40$ when $\alpha = 0.01$ this yields $k = 0.4$

Few spurious variables ever retained, yet 2^N possible models, namely 10^{12}.

Returning to our example:

- 16 irrelevant variables: the intercept and $z_{6,t} \ldots, z_{20,t}$.
- at $\alpha = 0.05$ we should retain 0.8 of a variable on average
- retention of irrelevant variables slightly higher at 3
- $\alpha = 0.01$ we should retain 0.16 of a variable on average – most of the time irrelevant variables would be eliminated

Explanations:

- Sampling variation – one draw from LDGP.
- 1-cut is only valid for perfectly orthogonal regressors. Although in population regressors are orthogonal, sample is correlated.

  PcGive → Other models → Descriptive statistics using PcGive → Means, standard deviations and correlations.

  Largest correlations ± 0.26.

  Need path search to ensure ordering doesn't matter (return to this shortly).

Consider the power of a t-test to retain relevant variables.

Denote the t-test as $\mathsf{t}\left(n,\psi\right)$ where n is the degrees of freedom and ψ is the non-centrality parameter, which is 0 under the null.

H_0: $\beta_i = 0$

To calculate the power to reject the null when $E\left[\mathrm{t}\right] = \psi > 0$:

$$P\left(t \geq c_\alpha | E\left[t\right] = \psi\right) \approx P\left(t - \psi \geq c_\alpha - \psi | H_0\right).$$

Approximate power if coefficient null **only tested once**:

t-test powers

ψ	α	$\mathsf{P}(\lvert \mathsf{t}\rvert \geq c_\alpha)$	$\mathsf{P}(\lvert \mathsf{t}\rvert \geq c_\alpha)^4$
1	0.05	0.16	0.001
2	0.05	0.50	0.063
2	0.01	0.26	0.005
3	0.05	0.85	0.512
3	0.01	0.64	0.168
4	0.05	0.98	0.902
4	0.01	0.91	0.686
6	0.01	1.00	0.997

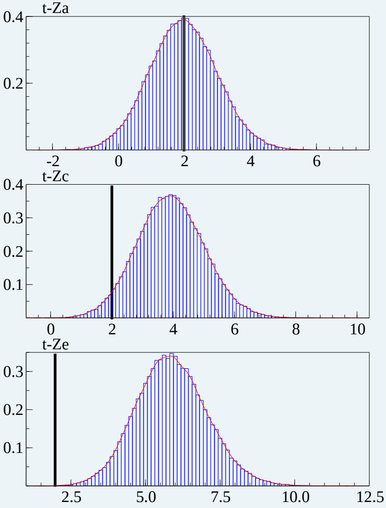
Low signal-noise variables will rarely be retained using t-tests when the null is tested.

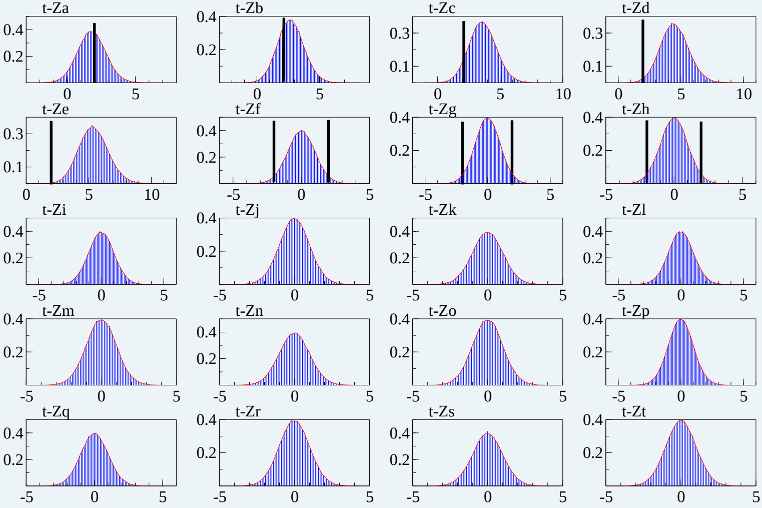
Therefore the key problem: retaining relevant variables.

Returning to our example:

- 5 relevant variables with non-centralities of 2,3,4,5 and 6
- at $\alpha = 0.05$ the probability of retaining z_1 is 0.51; z_2 is 0.85; z_3 is 0.98; z_4 is 1.00 and z_5 is 1.00.
  The probability of retaining all 5 variables is 0.42, to retain 4 it is 0.49, for 3 it is 0.08, for 2 it is 0.01, to retain 0 or 1 it is ≈ 0.
- at $\alpha = 0.01$ the probability of retaining z_1 is 0.23; z_2 is 0.65; z_3 is 0.92; z_4 is 0.99 and z_5 is 1.00.
  The probability of retaining all 5 variables is 0.14.
- Example keeps variables with $\psi = 3, 4, 5, 6$ but does not retain $\psi = 2$.

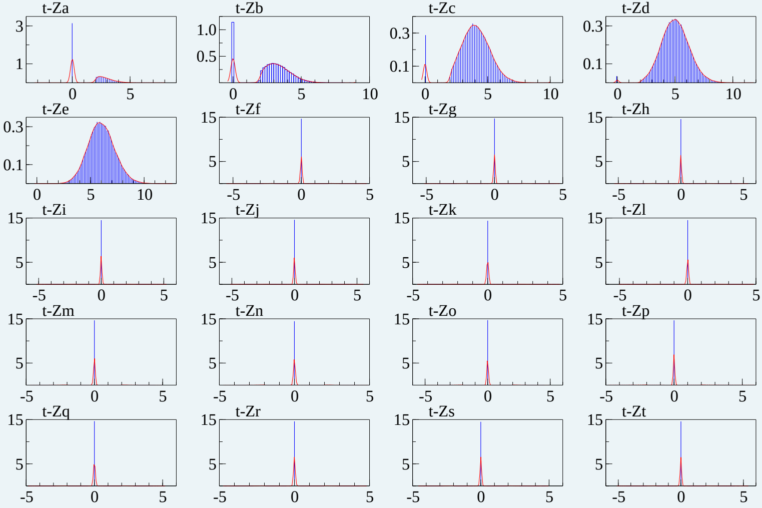
t-statistic distributions for DGP





t-Za
t-Zb
t-Zc
t-Zd
t-Ze
t-Zf
t-Zg
t-Zh
t-Zi
t-Zj
t-Zk
t-Zl
t-Zm
t-Zn
t-Zo
t-Zp
t-Zq
t-Zr
t-Zs
t-Zt

Unconditional t-statistic distributions for selected 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.

- $\mathsf{p}_{\alpha,i}^{dgp}$: probability of retaining i^{th} variable in DGP at size α.
- $1 - \mathsf{p}_{\alpha,i}^{dgp}$ is **cost of inference** (prob. of discarding relevant).
- M relevant, $m \leq M$ retained.
- $\mathsf{p}_{\alpha,i}^{gum}$: probability of retaining i^{th} variable in GUM.
- K irrelevant variables, $k \leq K$ retained.
- **Search costs** are $\sum_{i=1}^{M} \left(\mathsf{p}_{\alpha,i}^{dgp} - \mathsf{p}_{\alpha,i}^{gum} \right) + \sum_{j=1}^{K} \left(\mathsf{p}_{\alpha,j}^{gum} \right)$.

Return to example and estimate the DGP:

$$\begin{aligned} y_t &= \underset{(0.114)}{0.244} z_{1,t} + \underset{(0.107)}{0.384} z_{2,t} + \underset{(0.118)}{0.556} z_{3,t} + \underset{(0.11)}{0.323} z_{4,t} + \underset{(0.108)}{0.563} z_{5,t} \\ \sigma &= 1.086; \quad L = -147.6. \end{aligned} \tag{3}$$

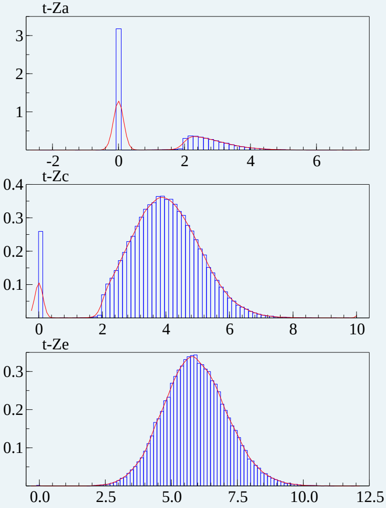
Using 1-cut rule at $\alpha = 0.05$, all relevant variables would be retained. Using 1-cut rule at $\alpha = 0.01$, $z_{1,t}$ would not be retained: a cost of inference.

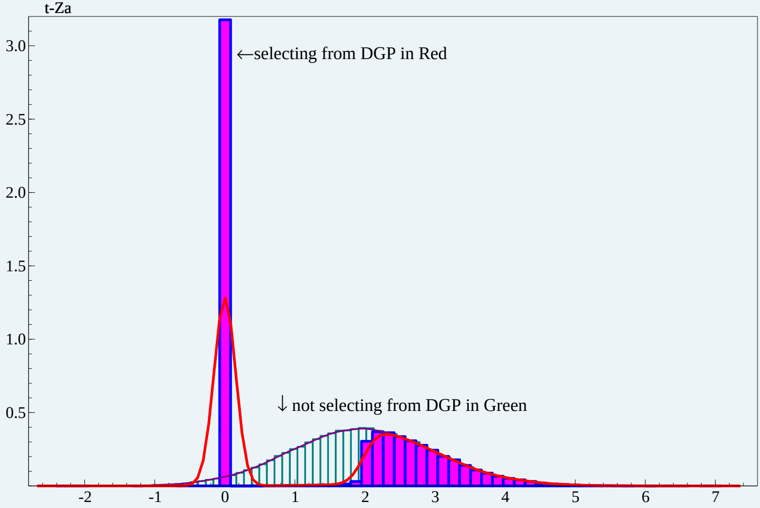
Compare to 1-cut rule commencing from GUM.
Search costs are:

- relevant variables that are retained in DGP but not in GUM – z_1 at 5%, none at 1%.
- retained irrelevant variables in GUM – z_6, z_7, z_{10} at 5%; z_7, z_{10} at 1%.

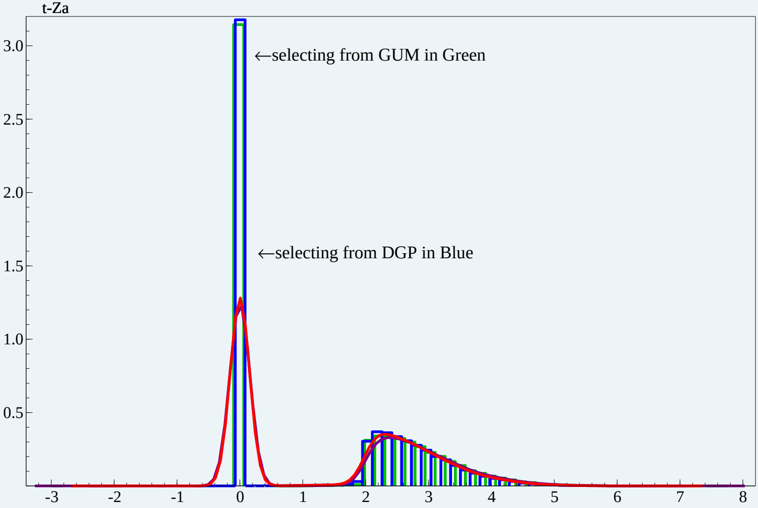
Simulations using *Autometrics* shows search costs can be small relative to costs of inference.

Unconditional t-statistic distributions for selected DGP




t-Za
3.0
2.5
2.0
1.5
1.0
0.5
←selecting from DGP in Red
↓ not selecting from DGP in Green
-2
-1
0
1
2
3
4
5
6
7

Selection from DGP or GUM



OXFORD MARTIN SCHOOL | UNIVERSITY OF OXFORD | Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL

How to find the source of Nile? Every path is explored; North, South, East and West, till success

Gets does this for model selection.

Search all reduction paths in general model

Path search gives impression of 'repeated testing'.
Confused with selecting from 2^N possible models
Our example just 21 variables, $2^{21} = 2097152$ possible models.
More realistic examples where $N > 100$: searching all possible models 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.

Does repeated testing distort selection?

1. Severe illness:
   more tests **increase** probability of **correct diagnosis**.
2. Mis-specification tests:
   if r independent tests τ_j conducted under null
   for small significance level η (critical value c_η):

$$\mathsf{P}(|\tau_j| < c_\eta \mid j = 1, \ldots, r) = (1 - \eta)^r \simeq 1 - r\eta.$$

   More tests **increase** probability of **false rejection**
   Suggests significance level η of 1% or tighter.
3. Repeated diagnostic tests: **probabilities unaltered**.

Conclude: no generic answer.

Specify GUM and select (e.g. at $\alpha = 0.05$):

$$
\begin{aligned}
y_t &= \underset{(0.114)}{0.269} + \underset{(0.104)}{0.391} z_{2,t} + \underset{(0.116)}{0.585} z_{3,t} + \underset{(0.108)}{0.322} z_{4,t} \\
&\quad + \underset{(0.106)}{0.573} z_{5,t} + \underset{(0.106)}{0.311} z_{7,t} + \underset{(0.11)}{0.275} z_{10,t} \qquad (4) \\
\sigma &= 1.051; \quad \mathsf{R}^2 = 0.488; \quad L = -143.24. \\
\mathsf{F}_{ar} &= 2.411; \quad \mathsf{F}_{arch} = 0.130; \quad \mathsf{F}_{hetero} = 0.974; \\
\mathsf{F}_{heteroX} &= 0.970; \quad \mathsf{F}_{reset} = 0.858; \quad \chi^2_{norm} = 0.551
\end{aligned}
$$

Compare across all results...

- Diagnostic checking
- Encompassing

Selection matters: only retain 'significant' variables.
Can correct final estimates for selection.

Convenient approximation that:

$$\mathsf{t}_{\widehat{\beta}} = \frac{\widehat{\beta}}{\widehat{\sigma}_{\widehat{\beta}}} \simeq \frac{\widehat{\beta}}{\sigma_{\widehat{\beta}}} \sim \mathsf{N}\left[\frac{\beta}{\sigma_{\widehat{\beta}}}, 1\right] = \mathsf{N}\left[\psi, 1\right]$$

when non-centrality of t-test is $\psi = \frac{\beta}{\sigma_{\widehat{\beta}}}$

Using Gaussian approximation:

$$\begin{aligned} \phi\left(w\right) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}w^2\right) \\ \Phi\left(w\right) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{w}\exp\left(-\frac{1}{2}x^2\right)\mathsf{d}x \end{aligned}$$

Doubly-truncated distribution–expected truncated t-value is:

$$\mathsf{E}\left[\left|\mathsf{t}_{\widehat{\beta}}\right| \mid \left|\mathsf{t}_{\widehat{\beta}}\right| > c_\alpha; \psi\right] = \psi^* \tag{5}$$

so observed |t|-value is unbiased estimator for ψ^*.
Thus, observe ψ^* when true non-centrality is ψ.
Sample selection induces:

$$\psi^* = \psi + \frac{\phi(c_\alpha - \psi) - \phi(-c_\alpha - \psi)}{1 - \Phi(c_\alpha - \psi) + \Phi(-c_\alpha - \psi)} = \psi + r\left(\psi, c_\alpha\right) \tag{6}$$

As know mapping $\psi^* \rightarrow \psi$, can correct by 'inversion':
$\psi = \psi^* - r\left(\psi, c_\alpha\right)$, albeit iteratively as r depends on ψ.

Applies as well to correcting $\widetilde{\beta}$ once ψ is known: for $\beta \geq 0$:

$$\mathsf{E}\left[\widetilde{\beta} \mid \widetilde{\beta} \geq \sigma_{\widetilde{\beta}} c_\alpha\right] = \beta\left(1 + \frac{r\left(\psi, c_\alpha\right)}{\psi}\right) = \beta\left(\frac{\psi^*}{\psi}\right) \tag{7}$$

Estimate ψ^* from $\mathsf{t}_{\widetilde{\beta}}$ then iteratively solve for ψ from (6):

$$\psi = \psi^* - r\left(\psi, c_\alpha\right) \tag{8}$$

so replace $r(\psi, c_\alpha)$ in (8) by $r(\mathsf{t}_{\widetilde{\beta}}, c_\alpha)$, and ψ^* by $\mathsf{t}_{\widetilde{\beta}}$:

$$\widetilde{\psi} = \mathsf{t}_{\widetilde{\beta}} - r\left(\mathsf{t}_{\widetilde{\beta}}, c_\alpha\right), \quad \text{then} \quad \widetilde{\widetilde{\psi}} = \mathsf{t}_{\widetilde{\beta}} - r\left(\widetilde{\psi}, c_\alpha\right) \tag{9}$$

leading to the bias-corrected parameter estimate:

$$\widetilde{\widetilde{\beta}} = \widetilde{\beta}\left(\widetilde{\widetilde{\psi}}/\mathsf{t}_{\widetilde{\beta}}\right). \tag{10}$$

from inverting (7).

Implementing bias correction

Bias corrects closely, not exactly, for relevant: over-corrects for some t-values (Hendry and Krolzig, 2005).

No impact on 'bias' of parameters of irrelevant variables as their $\beta_i = 0$, so unbiased with or without selection.

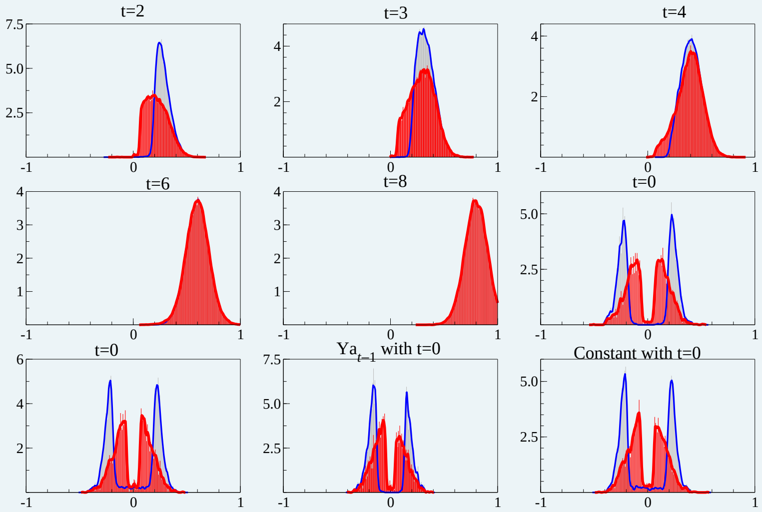
Some increase in MSEs of relevant variables
Correction exacerbates downward bias in unconditional estimates of relevant coefficients & increases MSEs slightly.

But remarkable decrease in MSEs of irrelevant variables.

First 'free lunch' of new approach:
obvious why in retrospect–most correction for $|\mathrm{t}|$ near c_α which holds for retained irrelevant variables.

Bias corrections applied to orthogonal variables:
two highly correlated regressors each have barely significant coefficients so large adjustment to both and hence their joint effect, if orthogonalized, only one would be adjusted.

OXFORD MARTIN SCHOOL UNIVERSITY OF OXFORD Institute for New Economic Thinking AT THE OXFORD MARTIN SCHOOL



Against such costs, bias correction considerably reduces MSEs of coefficients of retained irrelevant variables:
benefits both unconditional and conditional distributions.

Despite selecting from a very large set of potential variables:
nearly unbiased estimates of coefficients and
equation standard errors can be obtained;
little loss of efficiency from testing irrelevant variables,
some loss from not retaining relevant variables at large values of c_α;
huge gain by not commencing from an under-specified model.
Normal distribution has 'thin tails', so power loss from tighter significance levels rarely substantial,
but could be for fat-tailed error processes at tighter α.

Bias corrected coefficients in example

Bias correction code is not automated in OxMetrics but a simple Ox code can be applied.

Open: BiasCorrectionCode.ox

Paste in coefficient estimates and t-statistics.

Choose significance level.

Run to obtain 2-step corrected estimates:

$$\begin{aligned} y_t &= 0.162 + 0.380 z_{2,t} + 0.585 z_{3,t} + 0.278 z_{4,t} \\ &\quad + 0.573 z_{5,t} + 0.264 z_{7,t} + 0.188 z_{10,t} \\ \sigma &= 1.051; \quad \mathsf{R}^2 = 0.488; \quad L = -143.24. \end{aligned} \tag{11}$$

Hendry, D. F. and H.-M. Krolzig (2005).
The properties of automatic Gets modelling.
Economic Journal 115, C32–C61.