General-to-Specific Time Series Modelling

Felix Pretis

University of Oxford

Michaelmas 2017

Lecture 1: Model Discovery and Theory Embedding

General-to-Specific Time Series Modelling

- 4 Lectures (1.5hrs each)
  - Model Discovery and Theory Embedding
  - Indicator Saturation
  - Exogeneity
  - Theory of Reduction
- Centred around core papers with unifying theme:
  - Theories are incomplete (& likely wrong)
  - not imposing theory on data
  - via model selection from a general specification tackle empirical problems jointly and learn from the data
  - while retaining available theory insights

General-to-Specific Time Series Modelling

- 4 Lectures (1.5hrs each)
  - Model Discovery and Theory Embedding
  - Indicator Saturation
  - Exogeneity
  - Theory of Reduction
- Centred around core papers with unifying theme:
  - Theories are incomplete (& likely wrong)
  - not imposing theory on data
  - via model selection from a general specification tackle empirical problems jointly and learn from the data
  - while retaining available theory insights
- Slides online at: `www.felixpretis.org/teaching`
- Related exam questions: 2011Q7, 2012Q7, 2013Q3, 2014Q2, 2015Q2, 2016Q1, 2017Q1
  - no forecasting this course.
  - instead focus on exogeneity (only in past paper 2017Q1).

Core References for Lecture 1:

- Hendry and Johansen (2015)* – Embedding Theory
- Hendry and Krolzig (2005)* – GETS Modelling
- Hendry and Doornik (2014) – Model Discovery (book)

"The basis of econometrics, the economic theories that we had been led to believe in by our forefathers, were perhaps not good enough. It is quite obvious that if the theories we build to simulate actual economic life are not sufficiently realistic, that is, if the data we get to work on in practice are not produced the way that economic theories suggest, then it is rather meaningless to confront actual observations with relations that describe something else."

Haavelmo (1989), Nobel Lecture

Postulate:

$$y_t = \beta' x_t + \epsilon_t, \quad t = 1, \ldots, T \tag{1}$$

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

Many tests to '**discover**' departures from assumptions of (1), followed by recipes for 'fixing' them–
covert and unstructured empirical model discovery.

Same start (1), but aim to find a '**robust**' estimate of a **constant** β by selecting over $\mathcal{T}$, given **correct** set of relevant variables $\mathbf{x}$.

Worry about data contamination and outliers, so select sample, $\mathcal{T}^*$, where outliers least in evidence.

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

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

Same start (1), but aim to find a '**robust**' estimate of a **constant** β by selecting over $\mathcal{T}$, given **correct** set of relevant variables $\mathbf{x}$.

Worry about data contamination and outliers, so select sample, $\mathcal{T}^*$, where outliers least in evidence.

All 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 or distribution, assuming **correct** $\mathbf{x}$, **no data contamination**, **constant** β, etc., all rarely checked.

Each assumes away what the other approaches tackle.

Need to tackle them all jointly.

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

Starting from T observations on $\mathrm{N} > \mathrm{n}$ variables $\{\mathbf{x}_t\}$, 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 $\{\mathbf{1}_{\{t=t_i\}}\}$ – indicators for shifts, outliers etc.

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

Starting from T observations on $\mathsf{N} > \mathfrak{n}$ variables $\{\mathbf{x_t}\}$, aim to find $\boldsymbol{\beta}^*$ for s lagged functions $\mathbf{g}(\mathbf{x}^*_{\mathbf{t}}) \dots \mathbf{g}(\mathbf{x}^*_{\mathbf{t}-s})$ of a subset of $\mathfrak{n}$ variables $\mathbf{x}^*$, jointly with $\mathcal{T}^*$ and $\{\mathbf{1}_{\{\mathbf{t}=\mathbf{t}_i\}}\}$ – indicators for shifts, outliers etc.

Embed initial economic analysis $\mathbf{y} = \mathbf{f}(\mathbf{x})$ in a much more general empirical model.

Approach explained in Castle, Doornik, and Hendry (2011) extensive discussion in Hendry and Doornik (2014).

Seven stages for discovery in econometrics

1. **theoretical derivation** of the relevant set $\mathbf{x}$.
2. going **outside** current view by automatic creation of a general model from $\mathbf{x}$ embedding $\mathbf{y} = \mathbf{f}(\mathbf{x})$.
3. **search** by automatic selection to find viable representations – too large for manual labor.
4. criteria to **recognize** when search is completed – congruent parsimonious-encompassing model.
5. **quantification** of the outcome – translated into unbiasedly estimating the resulting model.
6. **evaluate** discovery to check its 'reality: new data, new tests or new procedures. Can also evaluate the selection process itself.
7. **summarize** vast information set in parsimonious but undominated model.

Approach is **not** atheoretic.

Theory formulations should be embedded in starting point (general unrestricted model – GUM), and can be retained without selection.

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

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

Extension of 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.

The Local DGP (LDGP) is the target for model selection

The set of variables $\{\mathbf{x_t}\}$ chosen for analysis will depend on the subject-matter theory, institutional knowledge, and previous evidence, so any theory-model **object** is directly related to the **target** LDGP.

But the LDGP is always unknown in practice, which is why Hendry and Doornik (2014) emphasize the need to **discover** the LDGP from the available evidence.

The Local DGP (LDGP) is the target for model selection

The set of variables $\{\mathbf{x_t}\}$ chosen for analysis will depend on the subject-matter theory, institutional knowledge, and previous evidence, so any theory-model **object** is directly related to the **target** LDGP.

But the LDGP is always unknown in practice, which is why Hendry and Doornik (2014) emphasize the need to **discover** the LDGP from the available evidence.

Doing so requires:

- formulating the theoretical framework;
- nesting that LDGP in a suitably general unrestricted model;
- while also embedding the theory model in that GUM;
- searching for the simplest acceptable representation;
- then stringently evaluating that selection for congruence, encompassing and invariance.

Fulfills all the steps for empirical discovery with theory evaluation.

Hendry and Johansen (2015): Embedding Theory
– an example of pitfalls of excluding relevant variables & non-stationarity:

Relevant variables $\mathbf{w_t}$ excluded from model of $\mathbf{y_t}$, mean-shifts in included policy variable alters outcome:

DGP for $\mathbf{y_t}$:

$$\mathbf{y_t} = \boldsymbol{\beta}'\mathbf{x_t} + \boldsymbol{\gamma}'\mathbf{w_t} + \boldsymbol{\epsilon_t} \tag{2}$$

where $\boldsymbol{\epsilon_t} \sim \mathsf{IN}[0, \sigma^2_{\epsilon}]$ and $\boldsymbol{\beta} \neq \mathbf{0}, \boldsymbol{\gamma} \neq \mathbf{0}$.

$$\mathbf{w_t} = \boldsymbol{\psi} + \boldsymbol{\Psi}\mathbf{x_t} + \mathbf{v_t} \tag{3}$$

where $\mathsf{E}[\mathbf{x_t}\mathbf{v_t}'] = \mathbf{0}$ and $\boldsymbol{\Psi} \neq \mathbf{0}$. Let $\mathsf{E}[\mathbf{x_t}] = \boldsymbol{\delta_1}$.

When relevant variables $\mathbf{w_t}$ are excluded, $\mathbf{y_t}$ given $\mathbf{x_t}$ becomes:

$$\mathbf{y_t} = \boldsymbol{\gamma}'\boldsymbol{\psi} + (\boldsymbol{\beta}' + \boldsymbol{\gamma}'\boldsymbol{\Psi})\boldsymbol{\delta_1} + (\boldsymbol{\beta}' + \boldsymbol{\gamma}'\boldsymbol{\Psi})(\mathbf{x_t} - \boldsymbol{\delta_1}) + \boldsymbol{\gamma}'\mathbf{v_t} + \boldsymbol{\epsilon_t}$$

where means are separated out such that $\mathsf{E}[\mathbf{x_t} - \boldsymbol{\delta_1}] = \mathbf{0}$ and hence $\mathsf{E}[\mathbf{y_t}] = \boldsymbol{\gamma}'\boldsymbol{\psi} + (\boldsymbol{\beta}' + \boldsymbol{\gamma}'\boldsymbol{\Psi})\boldsymbol{\delta_1}$.

The mis-specified regression model is then:

$$\mathbf{y_t} = \lambda_0 + \boldsymbol{\lambda}_1'\mathbf{x_t} + \mathbf{e_t} \tag{4}$$

and matches the LDGP (local DGP) with $\lambda_0 = \boldsymbol{\gamma}'\boldsymbol{\psi}$ and $\boldsymbol{\lambda_1} = \boldsymbol{\beta} + \boldsymbol{\Psi}'\boldsymbol{\gamma}$.

Let $\mathbf{x_t}$ be policy variables, where changes in their mean alter $\mathrm{E}[\mathbf{x_T}] = \boldsymbol{\delta_1}$ to $\mathrm{E}[\mathbf{x_{T+1}}] = \boldsymbol{\delta_2}$

Actual outcome would be an average change in y of:

$$\mathrm{E}[\mathrm{y}_{T+1}] - \mathrm{E}[\mathrm{y}_T] = \boldsymbol{\beta}'(\boldsymbol{\delta_2} - \boldsymbol{\delta_1}) \tag{5}$$

Let E_M denote the expectations operator based on the mis-specified model, a shift in $\mathbf{x}$ produces an average **anticipated** change of:

$$\mathrm{E}_M[\mathrm{y}_{T+1}] - \mathrm{E}_M[\mathrm{y}_T] = \boldsymbol{\lambda}_1'(\boldsymbol{\delta_2} - \boldsymbol{\delta_1}) = (\boldsymbol{\beta}' + \boldsymbol{\gamma}'\boldsymbol{\Psi})(\boldsymbol{\delta_2} - \boldsymbol{\delta_1}) \tag{6}$$

Resulting in an unexpected location shift of $\boldsymbol{\gamma}'\boldsymbol{\Psi}(\boldsymbol{\delta_2} - \boldsymbol{\delta_1})$ – could lead to adverse policy effect.

Large risks of under-specified models.

Four possible economic theory outcomes

1. **Theory exactly correct**:
   all aspects significant with anticipated signs, no other variables kept.
2. **Theory only part of explanation**:
   all aspects significant with anticipated signs, but other variables also kept as substantively relevant.
3. **Theory partially correct**:
   only some aspects significant with anticipated signs, and other variables also kept as substantively relevant.
4. **Theory not correct**:
   no aspects significant and other variables do all explanation.

Consider a theory model which correctly matches the data-generating process (DGP) by specifying that:

$$y_t = \beta' x_t + \epsilon_t \tag{7}$$

where $\epsilon_t \sim \text{IID}[0, \sigma_\epsilon^2]$ over $t = 1, \ldots, T$, and ϵ_t is independent of the m strongly exogenous variables $\{x_1, \ldots, x_t\}$, assumed to satisfy:

$$T^{-1} \sum_{t=1}^{T} x_t x_t' \xrightarrow{P} \Sigma_{xx}$$

which is positive definite, and:

$$T^{1/2} \left(\widehat{\beta} - \beta_0 \right) = \left(T^{-1} \sum_{t=1}^{T} x_t x_t' \right)^{-1} T^{-1/2} \sum_{t=1}^{T} x_t \epsilon_t$$
$$\xrightarrow{D} N_m \left[0, \sigma_\epsilon^2 \Sigma_{xx}^{-1} \right] \tag{8}$$

where β_0 is the constant population parameter.

Consider additional set of k exogenous variables $\mathbf{w_t}$ may also influence y_t, so postulate the more general model:

$$y_t = \beta' \mathbf{x_t} + \gamma' \mathbf{w_t} + \epsilon_t \tag{9}$$

although pop. parameter: $\gamma_0 = 0$ (because theory exactly correct).

$\mathbf{w_t}$ can be: known to be exogenous, functions of those, lagged variables, non-linear, and indicators for outliers or breaks.

Properties of estimators when embedding (correct) theory in larger model?

$\mathbf{x_t}$ and $\mathbf{w_t}$ can be orthogonalized by first computing:

$$\widehat{\mathbf{\Gamma}} = \left(\sum_{t=1}^{T} \mathbf{w_t} \mathbf{x_t'} \right) \left(\sum_{t=1}^{T} \mathbf{x_t} \mathbf{x_t'} \right)^{-1} \tag{10}$$

and defining the residuals $\widehat{\mathbf{u}}_t$ by:

$$\mathbf{w_t} = \widehat{\mathbf{\Gamma}} \mathbf{x_t} + \widehat{\mathbf{u}}_t \tag{11}$$

so that:

$$\sum_{t=1}^{T} \mathbf{x_t} \widehat{\mathbf{u}}_t' = \mathbf{0} \tag{12}$$

$\mathbf{x_t}$ and $\mathbf{w_t}$ can be orthogonalized by first computing:

$$\widehat{\mathbf{\Gamma}} = \left(\sum_{t=1}^{T} \mathbf{w_t x_t'} \right) \left(\sum_{t=1}^{T} \mathbf{x_t x_t'} \right)^{-1} \tag{10}$$

and defining the residuals $\widehat{\mathbf{u}}_t$ by:

$$\mathbf{w_t} = \widehat{\mathbf{\Gamma}} \mathbf{x_t} + \widehat{\mathbf{u}}_t \tag{11}$$

so that:

$$\sum_{t=1}^{T} \mathbf{x_t} \widehat{\mathbf{u}}_t' = \mathbf{0} \tag{12}$$

Then substituting:

$$\begin{aligned} y_t &= \beta' \mathbf{x_t} + \gamma' \mathbf{w_t} + \epsilon_t = \beta' \mathbf{x_t} + \gamma' \left(\widehat{\mathbf{\Gamma}} \mathbf{x_t} + \widehat{\mathbf{u}}_t \right) + \epsilon_t \\ &= \beta_+' \mathbf{x_t} + \gamma' \widehat{\mathbf{u}}_t + \epsilon_t, \end{aligned} \tag{13}$$

where $\beta_+ = \beta + \widehat{\mathbf{\Gamma}}' \gamma$. Note that $\beta_{0+} = \beta_0$ because $\gamma_0 = 0$.

Consequently, as DGP is $y_t = \beta_0' x_t + \epsilon_t$:

$$
\begin{aligned}
& T^{1/2} \begin{pmatrix} \widetilde{\beta}_+ - \beta_0 \\ \widetilde{\gamma} \end{pmatrix} \\
&= \begin{pmatrix} T^{-1} \sum_{t=1}^{T} x_t x_t' & T^{-1} \sum_{t=1}^{T} x_t \widehat{u}_t' \\ T^{-1} \sum_{t=1}^{T} \widehat{u}_t x_t' & T^{-1} \sum_{t=1}^{T} \widehat{u}_t \widehat{u}_t' \end{pmatrix}^{-1} \begin{pmatrix} T^{-1/2} \sum_{t=1}^{T} x_t \epsilon_t \\ T^{-1/2} \sum_{t=1}^{T} \widehat{u}_t \epsilon_t \end{pmatrix} \\
&= \begin{pmatrix} \left(T^{-1} \sum_{t=1}^{T} x_t x_t' \right)^{-1} T^{-1/2} \sum_{t=1}^{T} x_t \epsilon_t \\ \left(T^{-1} \sum_{t=1}^{T} \widehat{u}_t \widehat{u}_t' \right)^{-1} T^{-1/2} \sum_{t=1}^{T} \widehat{u}_t \epsilon_t \end{pmatrix} \\
&\xrightarrow{D} N_{m+k} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} \Sigma_{xx}^{-1} & 0 \\ 0 & \Sigma_{ww|x}^{-1} \end{pmatrix} \right] \qquad (14)
\end{aligned}
$$

Consequently, as DGP is $y_t = \beta_0' x_t + \epsilon_t$:

$$
\begin{aligned}
& T^{1/2} \begin{pmatrix} \widetilde{\beta}_+ - \beta_0 \\ \widetilde{\gamma} \end{pmatrix} \\
& = \begin{pmatrix} T^{-1} \sum_{t=1}^{T} x_t x_t' & T^{-1} \sum_{t=1}^{T} x_t \widehat{u}_t' \\ T^{-1} \sum_{t=1}^{T} \widehat{u}_t x_t' & T^{-1} \sum_{t=1}^{T} \widehat{u}_t \widehat{u}_t' \end{pmatrix}^{-1} \begin{pmatrix} T^{-1/2} \sum_{t=1}^{T} x_t \epsilon_t \\ T^{-1/2} \sum_{t=1}^{T} \widehat{u}_t \epsilon_t \end{pmatrix} \\
& = \begin{pmatrix} \left(T^{-1} \sum_{t=1}^{T} x_t x_t' \right)^{-1} T^{-1/2} \sum_{t=1}^{T} x_t \epsilon_t \\ \left(T^{-1} \sum_{t=1}^{T} \widehat{u}_t \widehat{u}_t' \right)^{-1} T^{-1/2} \sum_{t=1}^{T} \widehat{u}_t \epsilon_t \end{pmatrix} \\
& \xrightarrow{D} N_{m+k} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} \Sigma_{xx}^{-1} & 0 \\ 0 & \Sigma_{ww|x}^{-1} \end{pmatrix} \right] \qquad (14)
\end{aligned}
$$

as $\sum_{t=1}^{T} x_t \hat{u}_t' = 0$, so distribution of $\widetilde{\beta_+}$ in (14) **identical** to that of $\widehat{\beta}$ in (8): **unaffected** by embedding in larger model.

Only 'costs' of selection (embedding theory in broader model) are:

- chance retentions of some $\widehat{\mathbf{u}}_t$ from selection; and
- impact on **estimated** distribution of $\widetilde{\boldsymbol{\beta}}_+$ through $\widetilde{\sigma}^2_\epsilon$.

First can be offset by tight α (level of significance of selection).

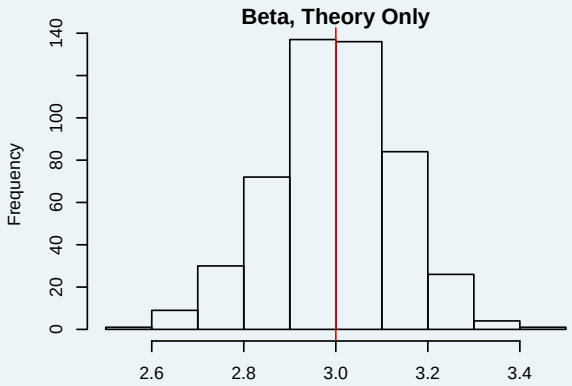
Under the null $\boldsymbol{\gamma_0 = 0}$ an unbiased estimate of σ^2_ϵ is:

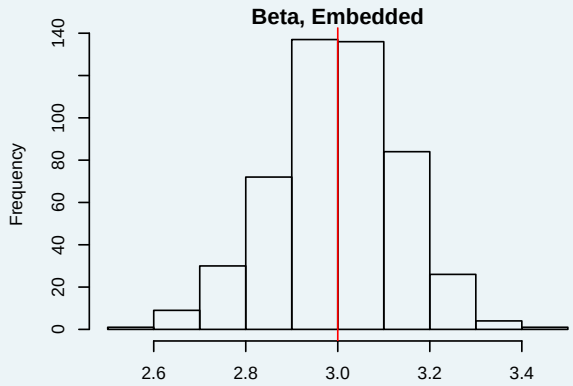
$$\widehat{\sigma}^2_\epsilon = (T - m)^{-1} \sum_{t=1}^{T} \left(y_t - \widetilde{\boldsymbol{\beta}}_+ \mathbf{x_t} \right)^2 \tag{15}$$

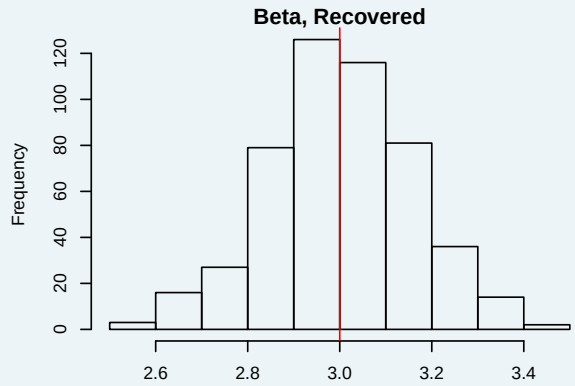
Theory Correct: Simulation Example

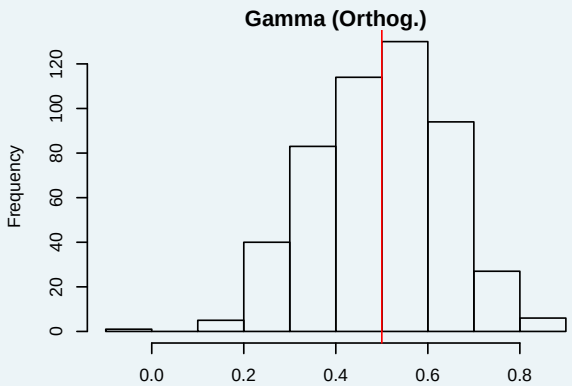
$$y_t = \beta' x_t + \gamma' w_t + \epsilon_t, \text{ and: } w_t = \Gamma x_t + \hat{u}_t \quad (16)$$

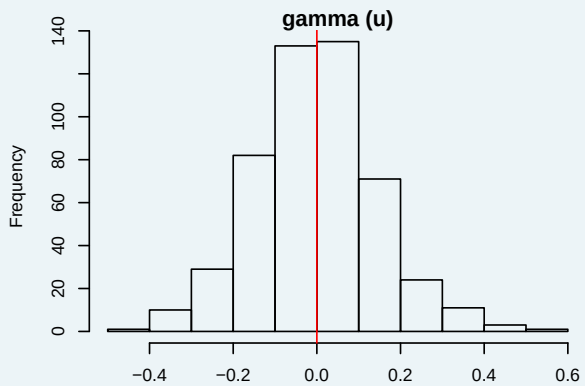
with: $\beta = 3, \gamma = 0, \Gamma = 0.5$, T $= 50$











Different when **theory model is only part of explanation**: defined as all aspects significant with anticipated signs, but other variables also kept as substantively relevant (some $\boldsymbol{\gamma_0 \neq 0}$).

Two distinct forms of under-specification:

1. omitting relevant functions or lags of variables in LDGP; avoided by sufficiently general initial model.
2. omitting relevant variables, $\mathbf{w_t}$, from the DGP; induces less useful LDGP–hard to avoid if $\mathbf{w_t}$ unknown.

Different when **theory model is only part of explanation**: defined as all aspects significant with anticipated signs, but other variables also kept as substantively relevant (some $\gamma_0 \neq 0$).

Two distinct forms of under-specification:

1. omitting relevant functions or lags of variables in LDGP; avoided by sufficiently general initial model.
2. omitting relevant variables, $\mathbf{w_t}$, from the DGP; induces less useful LDGP–hard to avoid if $\mathbf{w_t}$ unknown.

In DGP, $\gamma \neq 0$ coefficient on $\mathbf{x_t}$ is $\beta_0 + \gamma_0{}'\widehat{\Gamma}$.

- Selection can substantively improve the final model as able to retain some $\widehat{\mathbf{u}}_\mathbf{t}$
- Recover β by re-estimating non-orthogonalized.

Next, **when the theory is only partially correct**:
some aspects significant with anticipated signs,
but other aspects not significant (some $\boldsymbol{\beta_0 = 0}$), or 'wrong' signed,
with other variables also kept as substantively relevant (some $\boldsymbol{\gamma_0 \neq 0}$).

Next, **when the theory is only partially correct**:
some aspects significant with anticipated signs,
but other aspects not significant (some $\boldsymbol{\beta_0 = 0}$), or 'wrong' signed,
with other variables also kept as substantively relevant (some $\boldsymbol{\gamma_0 \neq 0}$).

Under alternative, $\boldsymbol{\gamma_0 \neq 0}$, model will result in biased, inefficient, possibly non-constant, estimates as:

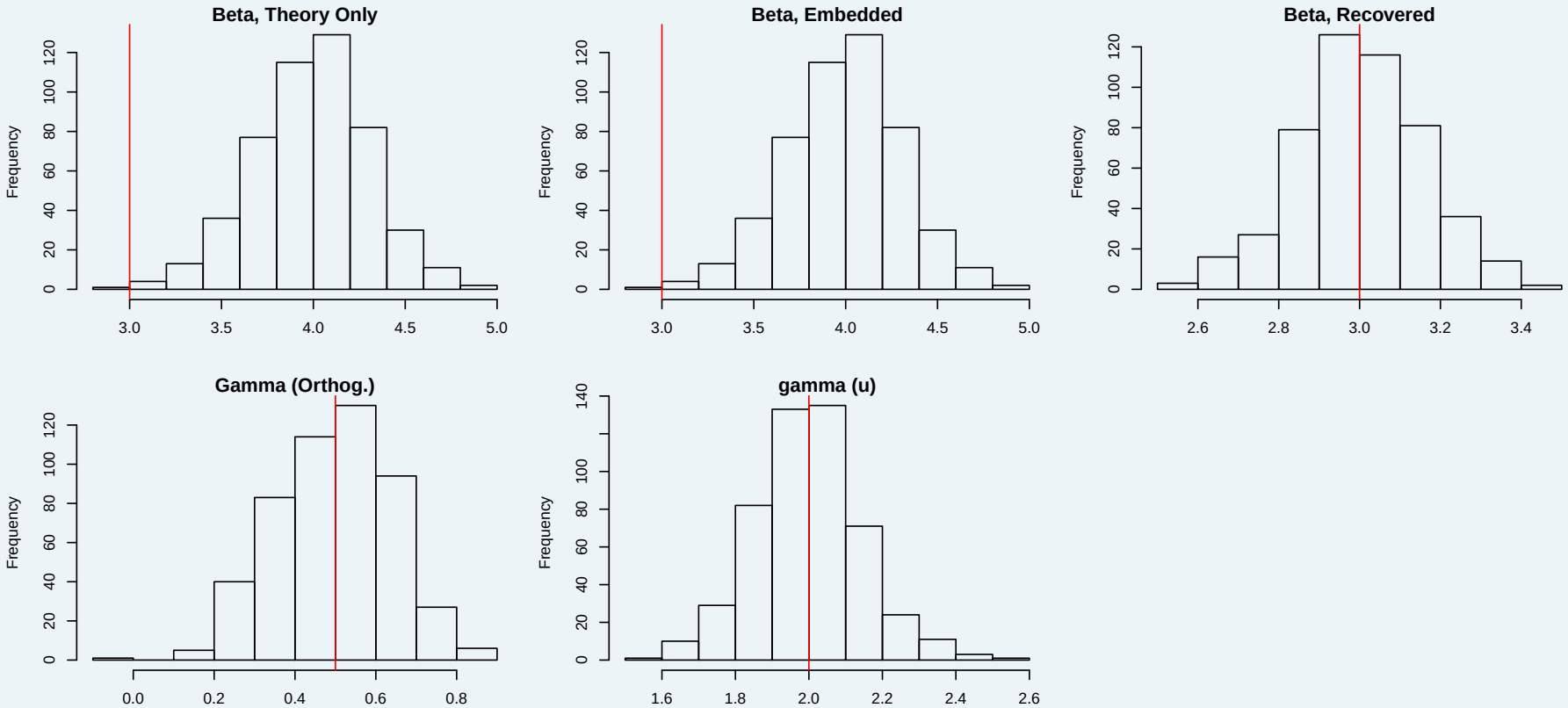
$$y_t = \beta' \mathbf{x}_t + \gamma' \left(\widehat{\Gamma} \mathbf{x}_t + \hat{\mathbf{u}}_t \right) + \epsilon_t = \left(\beta + \gamma' \widehat{\Gamma} \right)' \mathbf{x}_t + \gamma' \hat{\mathbf{u}}_t + \epsilon_t \quad (17)$$

Now forcing $\mathbf{x}_t$ when selecting from (17) will deliver an incorrect estimate of β, but some of the $\hat{\mathbf{u}}_t$ will be correctly retained, so an implied estimate of β can be derived from $\widetilde{\beta}_+ = \beta + \gamma' \widehat{\Gamma}$, $\widetilde{\gamma}$ and $\widehat{\Gamma}$.
A better estimate of $\widetilde{\sigma}^2_\epsilon$ should result.

Selection can also help when relevant variables, $\mathbf{w}_t$, omitted from LDGP and breaks occur.

$$y_t = \beta' x_t + \gamma' w_t + \epsilon_t, \text{ and: } w_t = \Gamma x_t + \hat{u}_t \quad (18)$$

with: $\beta = 3, \gamma = 2, \Gamma = 0.5$



Finally, **theory is now completely incorrect**:
no aspects significant and other variables do all explanation ($\boldsymbol{\beta_0 = 0}$).

Despite forcing $\mathbf{x_t}$ when $\boldsymbol{\beta_0 = 0}$, interpretation is awkward as coefficient of $\mathbf{x_t}$ is $\boldsymbol{\gamma' \widehat{\Gamma}}$.

Can be assessed using $\widetilde{\boldsymbol{\beta}}_+$, $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\Gamma}}$. Disastrous outcome if $\mathbf{w_t}$ omitted from initial model.

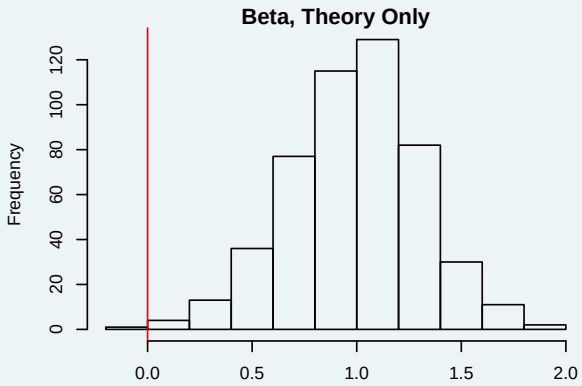
Win-win situation: theory kept if valid and complete;
yet learn when it is not correct –
empirical model discovery embedding theory evaluation.

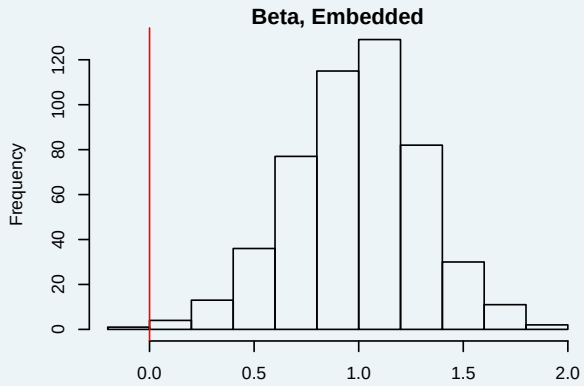
Interesting case is when $\mathrm{N} > \mathrm{T}$ for N candidates, so can automatic model selection work then?

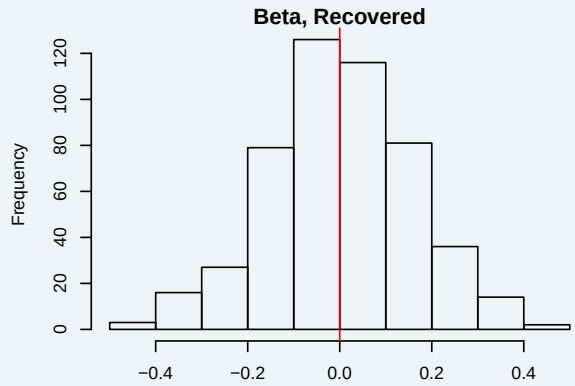
Theory not part: Simulation

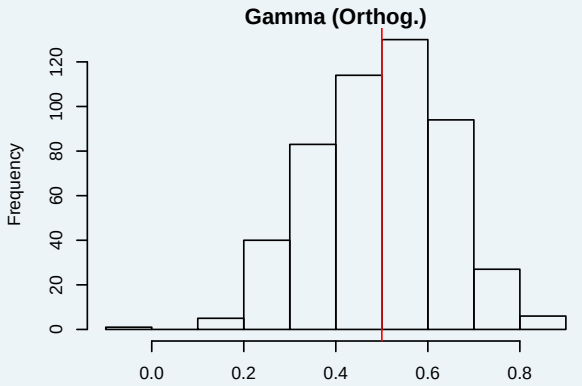
$$y_t = \beta' \mathbf{x_t} + \gamma' \mathbf{w_t} + \epsilon_t, \text{ and: } \mathbf{w_t} = \mathbf{\Gamma x_t} + \mathbf{\hat{u}_t} \quad (19)$$

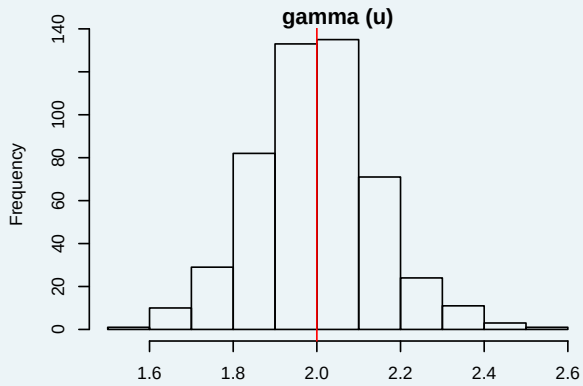
with: $\beta = 0, \gamma = 2, \Gamma = 0.5$











Theory-embedding: benefits of extended models

- almost costless when theory exactly correct
- benefits from learning from data
- also valid for endogenous regressors & IV (see appendix)

Still aim for parsimony: reduce general model (large set of $\mathbf{x_t}, \mathbf{w_t}$) to specific.

- Starting point: GUM (General Unrestricted Model)
- Target: LDGP (Local Data Generating Process)

Model selection to reach target.

Many methods for **model selection** (some frequently used but ineffective in realistic settings).

- Forward selection
- Step-wise regression
- 1-cut elimination
- Backward elimination
- Information criteria
- Lasso
- General-to-specific: Gets

Many methods for **model selection** (some frequently used but ineffective in realistic settings).

- Forward selection
- Step-wise regression
- **1-cut elimination**
- Backward elimination
- Information criteria
- Lasso
- **General-to-specific: Gets** (*Autometrics* in PcGive, *getsm* in R)

Performance of selection by IC well known for stationary and ergodic autoregressions:

- *AIC*, *SC* and *HQ* penalize log-likelihood by $f(N, T)$ for N parameters and sample T.
- *SC* (stricter) and *HQ* consistent:
  DGP$\subseteq$model selected with prob $\to 1$ as $T \to \infty$ relative to k
- Need to estimate all 2^N models to properly minimize information criterion.

$$\begin{aligned} SC &= \left(-2\widehat{\ell} + N \log T\right) T^{-1} \\ HQ &= \left(-2\widehat{\ell} + 2N \log\log T\right) T^{-1} \\ AIC &= \left(-2\widehat{\ell} + 2N\right) T^{-1} \end{aligned}$$

Problems with Information Criteria (IC):

- IC do not ensure adequate initial model specification (GETS tests GUM for congruency)
- Selection criteria too loose as $N \to T$
- Unclear how to use when $N >> T$
- 2^N becomes 'too large' very quickly

General-to-specific attempts to correct some of these drawbacks.

Two costs of selection: costs of **inference** and **search**

- First inevitable if tests have non-zero null and non-unit rejection frequencies under alternative
  Applies even if commence from LDGP.
  Measure costs of inference by RMSE of selecting or conducting inference on LDGP (alleviated if theory forced)
- When a GUM nests the LDGP, additional costs of search: calculate by increase in RMSEs for relevant variables when starting from the GUM as against the LDGP, plus those for retained irrelevant variables

Two costs of selection:

- costs of **inference** (alleviated if theory forced), 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: starting at GUM relative to LDGP

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

One-Cut model selection

Consider a perfectly orthogonal regression model:

$$y_t = \sum_{i=1}^{N} \beta_i z_{i,t} + \epsilon_t \tag{20}$$

$E[z_{i,t} z_{j,t}] = \lambda_{i,i}$ 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: $\beta_j = 0$:

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

Cut-off m between included and excluded variables is:

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

Larger values retained: all others eliminated.
Only one decision needed even for $N \geqslant 1000$:
'goodness of fit' is never considered.

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

Total number of variables: $N = m + k > T$ (with theory variables $m << T$)

Selection in blocks:

- Divide variables into sub blocks (retaining theory in each)
- Select variables in each block at $\alpha = 1/N$ overall

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

event	probability	retain
$P\left(\left\vert t_i\right\vert < c_\alpha,\ \forall i = 1, \ldots N\right)$	$(1-\alpha)^N$	0
$P\left(\left\vert t_i\right\vert \geqslant c_\alpha \left\vert \left\vert t_j\right\vert < c_\alpha,\ \forall j \neq i\right.\right)$	$N\alpha(1-\alpha)^{N-1}$	1
$\vdots$	$\vdots$	$\vdots$
$P\left(\left\vert t_i\right\vert < c_\alpha \left\vert \left\vert t_j\right\vert \geqslant c_\alpha,\ \forall j \neq i\right.\right)$	$N\alpha^{(N-1)}(1-\alpha)$	$N-1$
$P\left(\left\vert t_i\right\vert \geqslant c_\alpha,\ \forall i = 1, \ldots N\right)$	α^N	N

Average number of null variables retained is:

$$k = \sum_{i=0}^{N} i \frac{N!}{i!\,(N-i)!} \alpha^i (1-\alpha)^{N-i} = N\alpha. \tag{21}$$

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

Few spurious variables ever retained

Keeping relevant:

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

Denote the t-test as $t(n, \psi)$ 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 $E[t] = \psi > 0$:

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

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

t-test powers

ψ	α	$P(\lvert t\rvert \geqslant c_\alpha)$	$P(\lvert t\rvert \geqslant c_\alpha)^4$
1	0.05	0.16	0.001
2	0.05	0.50	0.063
2	0.01	0.26	0.005
3	0.01	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

50–50 chance of retaining when $E[t^2] = 4$ for $c_\alpha = 2$
Only 6% chance of keeping **4** such variables

Does repeated testing distort selection?

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

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

  More tests **increase** probability of **false rejection**.
  Suggests significance level η of 1% or tighter.

Conclude: no generic answer.

Hendry and Krolzig (2005):
Selection matters: only retain 'significant' variables.
Can correct final estimates for selection given selection rule.

Convenient approximation that:

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

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

Using Gaussian approximation:

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

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

$$\mathsf{E}\left[|\mathrm{t}_{\widehat{\beta}}| \mid |\mathrm{t}_{\widehat{\beta}}| > \mathrm{c}_{\alpha}; \psi\right] = \psi^{*} \tag{22}$$

so observed $|\mathrm{t}|$-value is unbiased estimator for ψ^{*}. Thus, observe ψ^{*} when true non-centrality is ψ.

Sample selection induces:

$$\psi^{*} = \psi + \frac{\phi(\mathrm{c}_{\alpha} - \psi) - \phi(-\mathrm{c}_{\alpha} - \psi)}{1 - \Phi(\mathrm{c}_{\alpha} - \psi) + \Phi(-\mathrm{c}_{\alpha} - \psi)} = \psi + \mathrm{r}\left(\psi, \mathrm{c}_{\alpha}\right) \tag{23}$$

As know mapping $\psi^{*} \to \psi$, can correct by 'inversion': $\psi = \psi^{*} - \mathrm{r}\left(\psi, \mathrm{c}_{\alpha}\right)$, albeit iteratively as r depends on ψ.

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

$$\mathsf{E}\left[\widetilde{\beta} \mid \widetilde{\beta} \geqslant \sigma_{\widetilde{\beta}} \mathrm{c}_{\alpha}\right] = \beta\left(1 + \frac{\mathrm{r}\left(\psi, \mathrm{c}_{\alpha}\right)}{\psi}\right) = \beta\left(\frac{\psi^{*}}{\psi}\right) \tag{24}$$

Estimate ψ^* from $t_{\widetilde{\beta}}$ then iteratively solve for ψ from (23):

$$\psi = \psi^* - r\left(\psi, c_\alpha\right) \tag{25}$$

so replace $r(\psi, c_\alpha)$ in (25) by $r(t_{\widetilde{\beta}}, c_\alpha)$, and ψ^* by $t_{\widetilde{\beta}}$:

$$\widetilde{\psi} = t_{\widetilde{\beta}} - r\left(t_{\widetilde{\beta}}, c_\alpha\right), \quad \text{then} \quad \widetilde{\widetilde{\psi}} = t_{\widetilde{\beta}} - r\left(\widetilde{\psi}, c_\alpha\right) \tag{26}$$

leading to the bias-corrected parameter estimate:

$$\widetilde{\widetilde{\beta}} = \widetilde{\beta}\left(\widetilde{\widetilde{\psi}}/t_{\widetilde{\beta}}\right). \tag{27}$$

from inverting (24).

Now illustrate **1-cut** by simulating selection of 10 relevant from 1000 candidate variables.
DGP is given by:

$$y_t = \beta_1 z_{1,t} + \cdots + \beta_{10} z_{10,t} + \epsilon_t, \quad (28)$$

$$\mathbf{z}_t \sim \mathsf{IN}_{1000}[\mathbf{0}, \mathbf{\Omega}], \quad (29)$$

$$\epsilon_t \sim \mathsf{IN}[0, 1], \quad (30)$$

where $\mathbf{z}_t' = (z_{1,t}, \cdots, z_{1000,t})$.
Set $\mathbf{\Omega} = \mathbf{I}_{1000}$ for simplicity, keeping regressors fixed between experiments

- $T = 2000$ observations.
- DGP coefficients, β, and non-centralities, ψ, in table 1
- Also theoretical powers of t-tests on individual coefficients.

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}
β	0.06	0.08	0.09	0.11	0.13	0.14	0.16	0.17	0.19	0.21
ψ	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5
$P_{0.01}$	0.28	0.47	0.66	0.82	0.92	0.97	0.99	1.00	1.00	1.00
$P_{0.001}$	0.10	0.21	0.38	0.55	0.76	0.89	0.96	0.99	1.00	1.00

Table : Coefficients β_i, non-centralities ψ_i, theoretical retention probabilities, $P_{\alpha,i}$.

GUM contains all 1000 regressors and intercept:

$$y_t = \beta_0 + \beta_1 z_{1,t} + \cdots + \beta_{1000} z_{1000,t} + u_t, \quad t = 1, \ldots, 2000.$$

DGP has first $n = 10$ variables relevant,

so 991 **variables irrelevant in GUM** (with intercept).

Report outcomes for $\alpha = 1\%$ and 0.1%. $M = 1000$ replications, where $1(\cdot)$ is indicator.

'*gauge*' denotes empirical null retention frequency.

'*potency*' is average non-null retention frequency.

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

All retained variables significant at c_α by design in 1-cut.

But not necessarily the case with automated Gets.

Irrelevant variables may be retained because of:

(a) diagnostic checking when a variable is insignificant, but deletion makes a diagnostic test significant, or

(b) with encompassing, a variable can be individually insignificant, but not jointly with all variables deleted so far.

Simulation gauges and potencies recorded in table 2.

α	Gauge	Potency	Theory power
1%	1.01%	81%	81%
0.1%	0.10%	69%	68%

Table : Potency and gauge for 1-cut selection with 1000 variables.

Gauges not significantly different from nominal sizes α:
selection is not 'oversized' even with 1000 variables

Potencies close to average theory powers of 0.811 and 0.684.

Close match between theory and evidence even when selecting just 10 relevant regressors from 1000 variables.

Also report MSEs after model selection.

$\widehat{\beta}_{k,i}$ is OLS estimate of coefficient on $x_{k,t}$ in GUM for replication i.
$\widetilde{\beta}_{k,i}$ is OLS estimate after model selection
$\widetilde{\beta}_{k,i} = 0$ when $z_{k,t}$ not selected in final model.
Calculate following MSEs:

$$\begin{aligned}
\text{MSE}_k &= \frac{1}{M}\sum_{i=1}^{M}\left(\widehat{\beta}_{k,i} - \beta_k\right)^2, \\
\text{UMSE}_k &= \frac{1}{M}\sum_{i=1}^{M}\left(\widetilde{\beta}_{k,i} - \beta_k\right)^2, \\
\text{CMSE}_k &= \frac{\sum_{i=1}^{M}\left[\left(\widetilde{\beta}_{k,i} - \beta_k\right)^2 \cdot 1_{(\widetilde{\beta}_{k,i}\neq 0)}\right]}{\sum_{i=1}^{M} 1_{(\widetilde{\beta}_{k,i}\neq 0)}}, \quad \left(\beta_k^2 \text{ if } \sum_{i=1}^{M} 1_{(\widetilde{\beta}_{k,i}\neq 0)} = 0\right)
\end{aligned}$$

Unconditional MSE (UMSE) substitutes $\widetilde{\beta}_{k,i} = 0$ when a variable is not selected.
Conditional MSE (CMSE) is computed over retained variables only.

Simulation MSEs

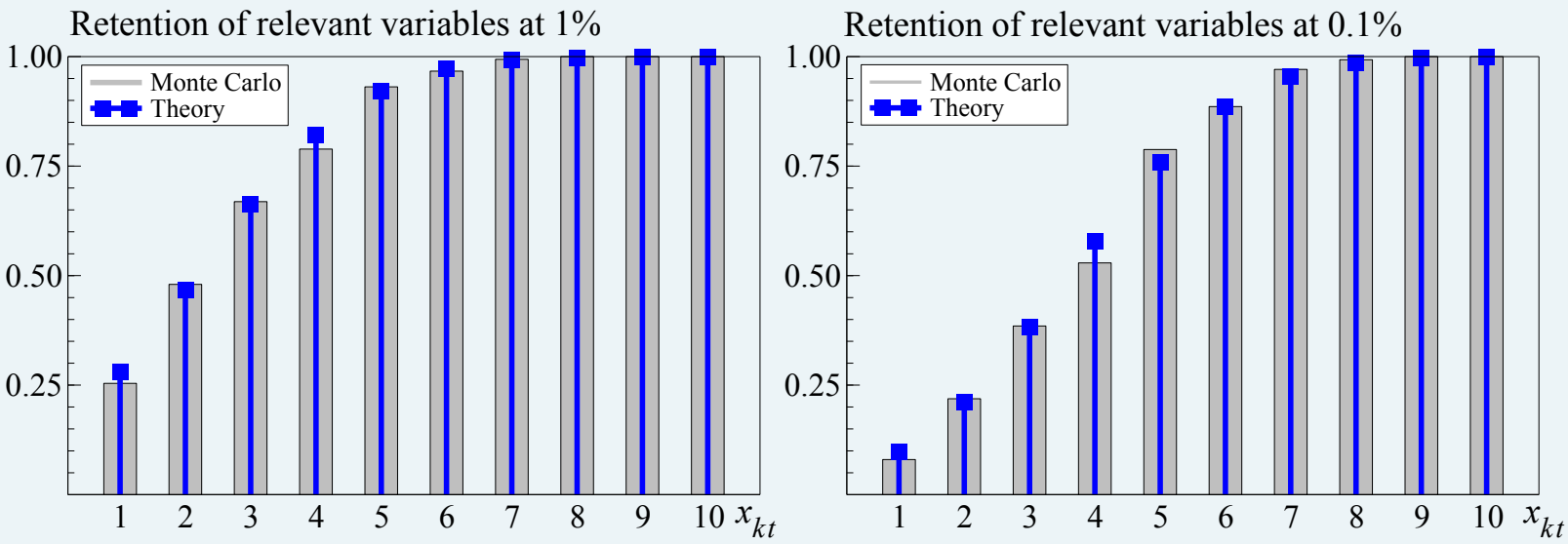


Figure shows that retention rates for individual relevant variables are as expected from the theory.

Consider impact of bias-correcting for selecting just those variables where $|t| \geqslant c_\alpha$

Impact of bias corrections on retained irrelevant and relevant variables, for $N = 1000$ and $n = 10$ in (20).

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
$\overline{\beta}$ after correction	0.38	0.60	1.2	1.3

Table : Average CMSEs, times 100, for 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.

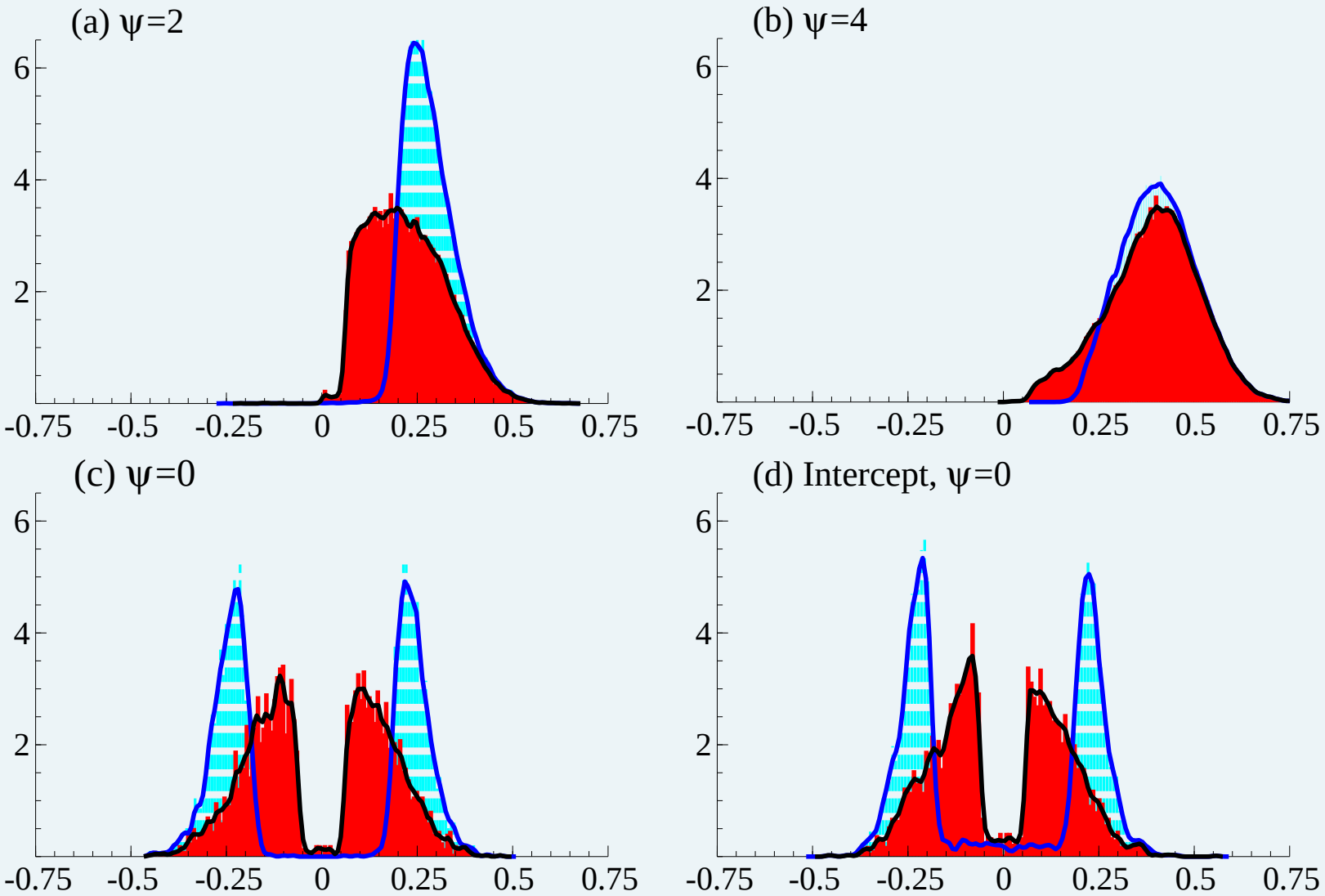
Coefficients of retained variables with $|t| \leqslant c_\alpha$ are not bias corrected–insignificant estimates set to zero.

- Bias corrects closely, not exactly, for relevant: over-corrects for some t-values.
- Some increase in MSEs of relevant variables.
  Correction exacerbates downward bias in unconditional estimates of relevant coefficients & increases MSEs slightly.
- No impact on 'bias' of estimated parameters of irrelevant variables as their $\beta_i = 0$, so unbiased with or without selection

But **remarkable decrease in MSEs of irrelevant variables**

First 'free lunch' of new approach.
Obvious why in retrospect–most correction for $|t|$ near c_α, which occurs for retained irrelevant variables.

Bias correcting cond. distrib. at 5%



DHSY (1978) re-visited:
UK consumption and income (quarterly: 1958:2 – 1976:2) with seasonals S_j.

Theory – permanent income hypothesis (PIH):

$$c_t = \underset{(0.07)}{0.60}\, c_{t-1} + \underset{(0.14)}{0.87} + \underset{(0.05)}{0.31}\, i_t - \underset{(0.01)}{0.12}\, S_1 - \underset{(0.005)}{0.01}\, S_2 - \underset{(0.003)}{0.03}\, S_3$$

```
AR 1-5 test:       F(5,66)    =    9.6825 [0.0000]**
ARCH 1-4 test:     F(4,69)    =    2.7946 [0.0327]*
Normality test:    Chi^2(2)   =    5.1375 [0.0766]
Hetero test:       F(7,69)    =    3.9719 [0.0011]**
RESET23 test:      F(2,69)    =   0.57147 [0.5673]
```

Lags (1-5), inflation $\Delta_4 p_t$, tax dummy D_t from DHSY. Orthogonalized w.r.t theory variables (').

$$
\begin{aligned}
c_t = & \underset{(0.07)}{0.60}\, c_{t-1} + \underset{(0.14)}{0.87} + \underset{(0.05)}{0.31}\, i_t - \underset{(0.01)}{0.12}\, S_1 - \underset{(0.005)}{0.01}\, S_2 - \underset{(0.003)}{0.03}\, S_3 \\
& - \underset{(0.09)}{0.04}\, c'_{t-2} + \underset{(0.09)}{0.05}\, c'_{t-3} + \underset{(0.09)}{0.72}\, c'_{t-4} - \underset{(0.12)}{0.02}\, c'_{t-5} \\
& + \underset{(0.05)}{0.15}\, i'_{t-1} - \underset{(0.05)}{0.03}\, i'_{t-2} + \underset{(0.04)}{0.04}\, i'_{t-3} - \underset{(0.05)}{0.09}\, i'_{t-4} - \underset{(0.05)}{0.19}\, i'_{t-5} \\
& - \underset{(0.08)}{0.33}\, \Delta_4 p'_t + \underset{(0.08)}{0.18}\, \Delta_4 p'_{t-1} + \underset{(0.007)}{0.002}\, D'_t
\end{aligned}
$$

```
AR 1-5 test:       F(5,46)    =   1.5562 [0.1914]
ARCH 1-4 test:     F(4,61)    =   2.9801 [0.0259]*
Normality test:    Chi^2(2)   = 0.080842 [0.9604]
Hetero test:       F(31,37)   =   1.4697 [0.1307]
RESET23 test:      F(2,49)    =   4.9168 [0.0113]*
```

Joint test of orthog.: $F(12, 51) = 13.114[0.0000]^{**}$

Model selection target size 1%, $N = 18$ (expect 0.01×18 spurious):

$$\begin{aligned} c_t &= \underset{(0.021)}{0.86}\, c_{t-4} - \underset{(0.0001)}{0.004}\, S_1 \\ &\quad + \underset{(0.03)}{0.25}\, i_t + \underset{(0.05)}{0.20}\, i_{t-1} - \underset{(0.03)}{0.31}\, i_{t-5} \\ &\quad - \underset{(0.08)}{0.33}\, \Delta_4 p_t + \underset{(0.07)}{0.26}\, \Delta_4 p_{t-1} + \underset{(0.002)}{0.008}\, D_t \end{aligned}$$

Can be written as equilibrium-correction, closely resembles DHSY. PIH but also allow for other effects that lie outside theory.

– Bias-correcting coefficients (technically valid for orthog. only):

$$\begin{aligned} c_t &= \underset{(0.021)}{0.86}\, c_{t-4} - \underset{(0.0001)}{0.002}\, S_1 \\ &\quad + \underset{(0.03)}{0.25}\, i_t + \underset{(0.05)}{0.20}\, i_{t-1} - \underset{(0.03)}{0.31}\, i_{t-5} \\ &\quad - \underset{(0.08)}{0.33}\, \Delta_4 p_t + \underset{(0.07)}{0.22}\, \Delta_4 p_{t-1} + \underset{(0.002)}{0.007}\, D_t \end{aligned}$$

Theory embedding and model discovery

- Theories incomplete & risks of under-specification
- Embedding theory in larger more general models
  - (Almost) costless when theory exactly correct
  - Learning from data under alternative
- Simplify general models to parsimonious specific ones
  - Easy to control false-positives
  - Challenge to retain relevant
  - Bias-correct for selection

Castle, J. L., J. A. Doornik, and D. F. Hendry (2011).
Evaluating automatic model selection.
Journal of Time Series Econometrics 3 (1), **DOI: 10.2202/1941–1928.1097.**

Davidson, J. E., D. F. Hendry, F. Srba, and S. Yeo (1978).
Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the united kingdom.
The Economic Journal, **661–692.**

Haavelmo, T. (1989).
Prize lecture.
Sveriges Riksbank: Prize in Economic Sciences in Memory of Alfred Nobel.

Hendry, D. F. and J. A. Doornik (2014).
Empirical Model Discovery and Theory Evaluation.
Cambridge, Mass.: MIT Press.

Hendry, D. F. and S. Johansen (2015).
Model discovery and Trygve Haavelmo's legacy.
Econometric Theory 31, **93–114.**

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

Appendix

- 'Theory embedding' with Endogenous regressors

Results generalize directly to instrumental variables

Some regressors not predetermined (endog.) and theory model is still:

$$y_t = \beta' x_t + \epsilon_t \tag{31}$$

where $\epsilon_t \sim \mathsf{IN}[0, \sigma^2_\epsilon]$, now ϵ is independent of the $m \geqslant n$ instrumental variables $z_1, \ldots, z_t$ where $(m + n < T)$.

The DGP if theory is correct has the form:

$$\mathbf{x_t} = \mathbf{\Pi z_t} + \boldsymbol{\zeta_t} \tag{32}$$

$$y_t = \boldsymbol{\beta}' \mathbf{\Pi z_t} + \boldsymbol{\eta_t} \tag{33}$$

where (η_t, ζ_t) are IID$[\mathbf{0}, \mathbf{\Omega}]$ with $\Omega = \begin{pmatrix} \sigma_\eta^2 & \sigma'_{\eta\zeta} \\ \sigma_{\zeta\eta} & \mathbf{\Omega}_\zeta \end{pmatrix}$ and $(\eta_t, \boldsymbol{\zeta_t})$

independent of $\mathbf{z_1}, \ldots, \mathbf{zt}$ but $\epsilon_t = y_t - \boldsymbol{\beta}' \mathbf{x_t} = \boldsymbol{\eta_t} - \boldsymbol{\beta}' \boldsymbol{\zeta_t}$ correlated with $\mathbf{x_t}$ as

$$\text{Cov}[\mathbf{x_t} \boldsymbol{\epsilon_t}] = \sigma_{\zeta\eta} - \mathbf{\Omega}_\zeta \beta \tag{34}$$

Instrumental variables estimation given by two-stage least squares:

$$\widehat{\beta} = \beta_0 + \left[\left(\sum_{t=1}^{T} x_t z_t' \right) \left(\sum_{t=1}^{T} z_t z_t' \right)^{-1} \left(\sum_{t=1}^{T} z_t x_t' \right) \right]^{-1} \times \left(\sum_{t=1}^{T} x_t z_t' \right) \left(\sum_{t=1}^{T} z_t z_t' \right)^{-1} \sum_{t=1}^{T} z_t \epsilon_t \tag{35}$$

so that:

$$T^{1/2} \left(\widehat{\beta} - \beta_0 \right) \xrightarrow{D} N_m \left[0, \sigma_\epsilon^2 Q^{-1} \right] \tag{36}$$

where we assume positive definite Q:

$$Q = \operatorname*{plim}_{T \to \infty} \left[\left(\frac{1}{T} \sum_{t=1}^{T} x_t z_t' \right) \left(\frac{1}{T} \sum_{t=1}^{T} z_t z_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} z_t x_t' \right) \right]$$

Let:

$$\widehat{\mathbf{\Pi}} = \left(\frac{1}{\mathsf{T}} \sum_{\mathbf{t}=1}^{\mathsf{T}} \mathbf{x_t} \mathbf{z}_\mathbf{t}' \right) \left(\frac{1}{\mathsf{T}} \sum_{\mathbf{t}=1}^{\mathsf{T}} \mathbf{z_t} \mathbf{z}_\mathbf{t}' \right)^{-1}$$

and define:

$$\widehat{\mathbf{x}}_\mathbf{t} = \widehat{\mathbf{\Pi}} \mathbf{z_t} \text{ with } \widehat{\xi}_\mathbf{t} = \mathbf{x_t} - \widehat{\mathbf{x}}_\mathbf{t} = \left(\mathbf{\Pi} - \widehat{\mathbf{\Pi}} \right) \mathbf{z_t} + \xi_\mathbf{t},$$

then a 2SLS reformulation that is algebraically convenient is:

$$\mathbf{y_t} = \beta' \widehat{\mathbf{x}}_\mathbf{t} + \mathbf{e_t} \tag{37}$$

where:

$$\mathbf{e_t} = \epsilon_\mathbf{t} + \beta' \widehat{\xi}_\mathbf{t} = \eta_\mathbf{t} + \beta' (\xi_\mathbf{t} - \widehat{\xi}_\mathbf{t})$$

so that:

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\widehat{\mathbf{x}}_t e_t = \operatorname*{plim}_{T\to\infty}\widehat{\mathbf{\Pi}}\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\left(\eta_t+\beta'\left(\xi_t-\widehat{\xi}_t\right)\right)=\mathbf{0}$$

Now include an additional set of k candidate exogenous variables $\mathbf{w_t}$ beyond theory:

$$\begin{aligned} y_t &= \beta'\mathbf{\Pi}\mathbf{z_t}+\gamma'\mathbf{w_t}+\eta_t \\ \mathbf{x_t} &= \mathbf{\Pi}\mathbf{z_t}+\xi_t \end{aligned} \tag{38}$$

where $\gamma_0=\mathbf{0}$, and the $\mathbf{x_t}$ are retained. Since $\gamma_0=\mathbf{0}$, when the $\widehat{\mathbf{x}}_t=\widehat{\mathbf{\Pi}}\mathbf{z_t}$ and $\mathbf{w_t}$ are orthogonalized as before:

$$\begin{aligned} y_t &= \beta'\widehat{\mathbf{x}}_t+\gamma'\mathbf{w_t}+\eta_t+\beta'\left(\xi_t-\widehat{\xi}_t\right) \\ &= \beta'\widehat{\mathbf{x}}_t+\gamma'\left(\widehat{\mathbf{\Gamma}}\widehat{\mathbf{x}}_t+\widehat{\mathbf{u}}_t\right)+\mathbf{e_t}=\beta'_+\widehat{\mathbf{x}}_t+\gamma'\widehat{\mathbf{u}}_t+\mathbf{e_t} \end{aligned} \tag{39}$$

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

When theory model is the DGP, by orthogonality:

$$
\begin{aligned}
&T^{1/2}\begin{pmatrix} \widetilde{\beta}_+ - \beta_0 \\ \widetilde{\gamma} \end{pmatrix} \\
&= \begin{pmatrix} T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{x}}_t\widehat{\mathbf{x}}_t' & T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{x}}_t\widehat{\mathbf{u}}_t' \\ T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{u}}_t\widehat{\mathbf{x}}_t' & T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{u}}_t\widehat{\mathbf{u}}_t' \end{pmatrix}^{-1} \begin{pmatrix} T^{-1/2}\sum_{t=1}^{T}\widehat{\mathbf{x}}_t e_t \\ T^{-1/2}\sum_{t=1}^{T}\widehat{\mathbf{u}}_t e_t \end{pmatrix} \\
&= \begin{pmatrix} \left(T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{x}}_t\widehat{\mathbf{x}}_t'\right)^{-1} T^{-1/2}\sum_{t=1}^{T}\widehat{\mathbf{x}}_t e_t \\ \left(T^{-1}\sum_{t=1}^{T}\widehat{\mathbf{u}}_t\widehat{\mathbf{u}}_t'\right)^{-1} T^{-1/2}\sum_{t=1}^{T}\widehat{\mathbf{u}}_t e_t \end{pmatrix} \\
&\xrightarrow{D} N_{m+k}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma_\eta^2 \begin{pmatrix} \boldsymbol{\Sigma}_{\widehat{\mathbf{x}}\widehat{\mathbf{x}}}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathbf{ww}|\mathbf{z}}^{-1} \end{pmatrix}\right] \qquad (40)
\end{aligned}
$$

Estimator $\widetilde{\beta}_+$ is again identical to the estimator $\widehat{\beta}$ in theory-only model, independently of the inclusion or exclusion of any or all of the $\widehat{\mathbf{u}}_t$.