Econometrics 871
Time Series

How to do econometrics properly:

Some thoughts on significance, data mining and model selection

Some questions

- **Should we care about statistical significance?**
- If so, how do we do it correctly?
  - When are these tests reliable as measures of “good models” or “strong results”?
    - The costs, risks, benefits and/or necessity of data mining

Should we care about statistical significance?

I assigned two suggested readings on this:

- McCloskey (1999) – a quite unusually emotional letter that is published as an academic contribution

- Hoover and Siegler (2008) – a considered, but also somewhat personal, response to McCloskey (and co-authors') universe of contributions.

There are many additional interesting parts to this literature, in particular the main two contributions of McCloskey and Ziliak that Hoover and Siegler respond to, and McCloskey's magnum opus:

McCloskey, D.N. and Ziliak, S.T., 1996. The standard error of regressions. *Journal of economic literature*, *34*(1), pp.97-114.

Ziliak, S.T. and McCloskey, D.N., 2004. Size matters: the standard error of regressions in the American Economic Review. *The Journal of Socio-Economics*, *33*(5), pp.527-546.

McCloskey, D.N. 1998. The Rhetoric of Economics (2nd ed.), Madison, WI: University of Wisconsin Press.

Deirdre McCloskey’s view

In my reading, the essential parts can be summarized as follows:

- Econometric publications often confuse *statistical significance* with *economic significance*
- A publishable paper typically needs statistically significant results to be accepted
- This has led to bad science: a lot of effort on establishing statistically significant results that do not contribute to scientifically important economic knowledge

Deirdre McCloskey’s view

Her example of this (conveniently, given our tutorials) is based on Purchasing Price Parity (PPP) which we expressed in logarithmic form in the cointegration setting as:

$$q_t = p_t - p_t^* - e_t$$

- Suppose the estimated coefficient on p_t^* is 0.999 with such a small standard error that the alternative hypothesis of 1 is rejected

- This is *statistical* evidence against the hypothesis, but is this sufficient *economic* evidence against the hypothesis?

- Taking into account measurement error, different baskets of goods, economic constraints to arbitrage, the conclusion she is aiming at is obviously correct. Economically, a 0.999 coefficient practically means that PPP holds

- Her argument is that many policy valuable coefficients are ignored because of lack of statistical significance and many unimportant economic effects are focussed on, just because they are statistically significant.

McCloskey and Ziliak

In two survey papers they analyse all* the contributions to the American Economic Review (one of the top journals in the world) that employ regression analysis

  - 182 papers in the 1980’s (McCloskey and Ziliak, 1996)
  - 137 papers in the 1990’s (McCloskey and Ziliak, 2004)
- Ask a battery of (19) questions, e.g.
  - Do the authors distinguish between statistical vs economic significance?
  - Do they evaluate the power of statistical tests?
  - Do they discuss the economic significance of the results?
- Some findings (there are many)
  - They find that the problem (as they see it) has become worse:
    - In the 1980s, 70%, and in the 1990s, 82%, did not distinguish between statistical / economic significance
  - 81% (1990s) of papers looked just for a “correct sign” and did not discuss the magnitude

Counter arguments to McCloskey and Co-authors

The negative responses to her claims are often very critical and excitingly aggressive.

These ones make strong logical points:

- Hoover and Siegler (2008)
    - Agree: economic and statistical significance are distinct. Reject: economists systematically mistake the two or ignore economic significance
    - Argue against the methods they employ:
        - Questions they ask are arbitrary, deeply subjective, some redundant leading to double-counting of “problems”
        - Can provide no mapping from the analysed text to the conclusion reached, no way to measure consistency across the two surveys – hence, this is not a replicable study and itself bad science.

Counter arguments to McCloskey and Co-authors

The negative responses to her claims are often very critical and excitingly aggressive.

These ones make strong logical points:

- Hoover and Siegler (2008)
  - Argues that significance testing is crucial
    - A measure of signal strength (signal-to-noise ratio)
    - Must be done correctly
    - An economically large but noisy signal must be treated with caution – it depends on the cost-benefits of committing either a Type 1 error (falsely rejecting a true null hypothesis) or a Type 2 error (mistakenly not rejecting a false null hypothesis).
    - Consider a lethal disease that can be cured with medication with unpleasant but non-fatal side-effects:
      H0: the patient has the disease
      - Type 1 error: falsely concluding that the patient does not have the disease leads to death
      - Type 2 error: falsely concluding that the patient does have the disease leads to unnecessary side-effects, but not death
    - If the disease has a low fatality rate, but the medication may also cause death, the trade-off changes. The same applies to any economic policy question, although it can be very difficult to specify the cost of a mistake
    - Shows that it is often used in exactly the same way in other sciences (contrary to McCloskey’s claims)

Counter arguments to McCloskey and Co-authors

- Spanos (2008) writes a review of their book that summarizes their main arguments

  Spanos, A., 2008. “Stephen T. Ziliak and Deirdre N. McCloskey's The cult of statistical significance: how the standard error costs us jobs, justice, and lives”. Ann Arbor (MI): The University of Michigan Press, 2008, xxiii+ 322 pp. *Erasmus Journal for Philosophy and Economics*, *1*(1), pp.154- 164.

- Deeply criticizes
  - Their reading of the history of statistics
  - Argues that some of their alternative are equivalent to significance testing
  - Their suggested approaches that side-steps a critical issue
    - One cannot argue for substantive results if a statistical model is not adequate, which requires specification testing.

My take:

- Positive: McCloskey presents many valid issues with the rhetoric of economics which may be problematic for the scientific value of our contributions
  - It is worth taking note of and evaluating her criticisms when you do econometrics
  - For those that you find convincing, make sure you don't do them
- Negative: Her approach reads as a little too rhetorical itself
  - The language, analyses and selections seem tailored to sell what she is offering
  - Colourful and engaging language is not problematic if she is right, but she seems to be doing exactly what she is accusing everyone of:
    - Picking the interpretation to suit the cause
    - Not providing a replicable statistical basis for her empirical claims
- I believe
  - Economic and statistical significance are *both* essential parts of any scientific contribution to economics
  - Statistical significance tests can be abused/misused in which case they are misleading and unscientific. Our next topic.

Some questions

- Should we care about statistical significance?
- **If so, how do we do it correctly?**
    - When are these tests reliable as measures of “good models” or “strong results”?
        - The costs, risks, benefits and/or necessity of data mining

Data Mining

- Definition from Hoover and Perez (2000):

  "'Data mining' refers to a broad class of activities that have in common, a search over different ways to process or package data statistically or econometrically with the purpose of making the final presentation meet certain design criteria."

- Spanos (2000) paraphrases Mayo (1996):

  "the data are being utilized for double-duty, to arrive at a claim (a model, an estimator, a test or some other inference proposition) in such a way that the claim is constrained to satisfy some criteria (e.g., fit) but the same data is regarded as supplying evidence in support of the claim arrived at."

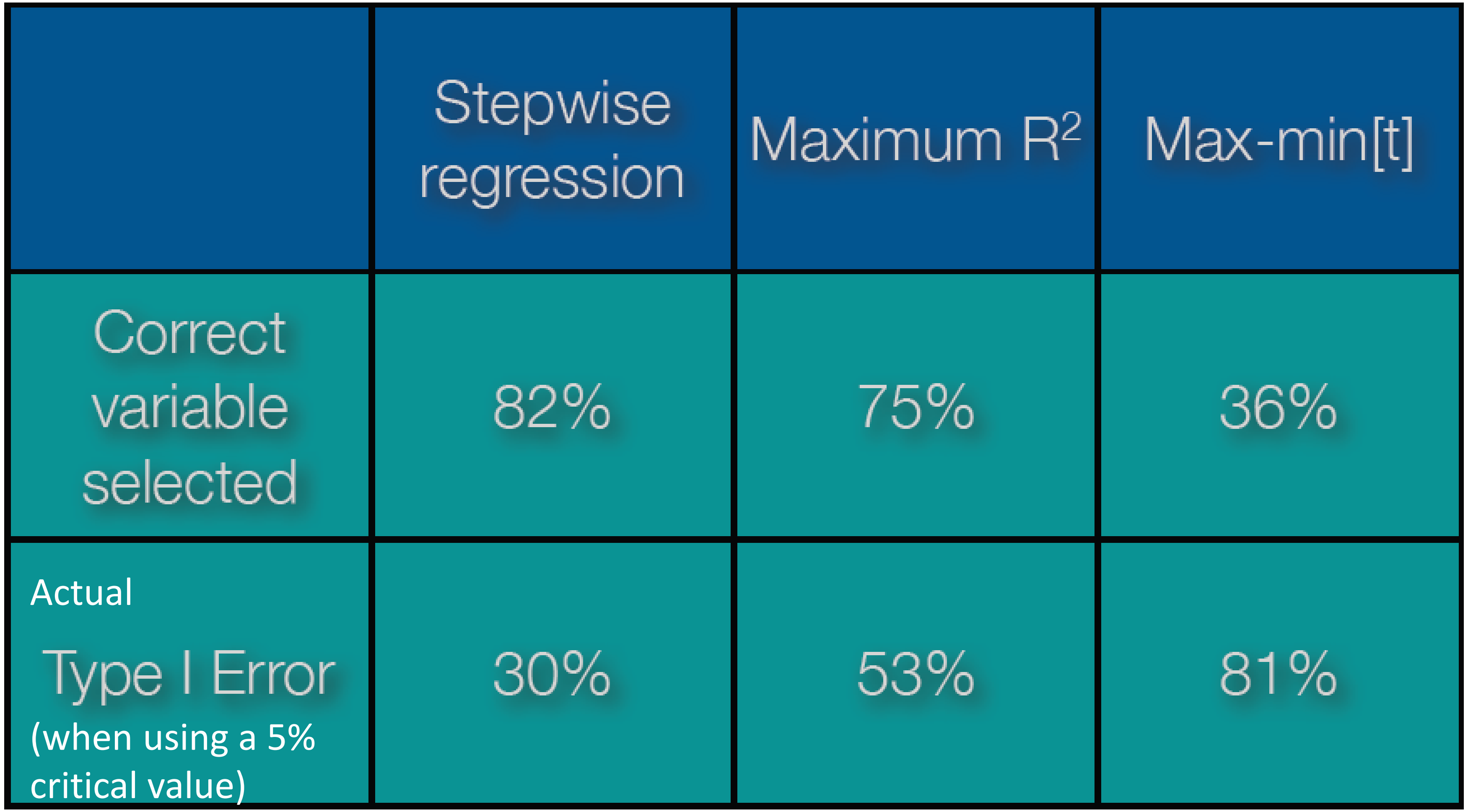
Data Mining

- Hoover and Perez (2000) identify three attitudes toward data mining:
  - It is bad: should be avoided, or if done, statistical tests should be adjusted to account for the repeated uses of the same data
  - It is bad but inevitable: given this, only robust models should be accepted (i.e. those that survive any data-mining attempt to destroy them)
  - It is essential: the only hope of uncovering the true DGP is by intelligently mining the data
- Spanos (2000) makes a similar distinction between
  - Warranted (unproblematic) data mining
  - Unwarranted (problematic) data mining

Data Mining: the bad

- Unwarranted data mining changes the true size of significance tests
  - Sampling variation may give spurious significant relationships
- Seminal contribution:
  - Lovell (1983) presents an early simulation study summarized in Hoover and Perez (2000)
  - They characterize Lovell (1983) as a study to show that data mining is always bad
  - This happens when the validity of a model is only based on statistical “goodness-of-fit”, regardless of the method by which it is achieved
- Lovell’s experiment
  - Key problem: selecting the correct variables that belong in the DGP
  - For a variety of simulated true DGP’s, ranging from
    - No relationship between dependent and explanatory variables, to
    - A true DGP with two explanatory variables
  - How do three model selection methods fare at uncovering the truth?
    - Step-wise regression – successively add explanatory variables
    - Max adjusted R-squared – select the best among all possible two variable regressions
    - Max-min t stat – select the model where the smallest of two t-stats is the largest across all the regressions

Results:



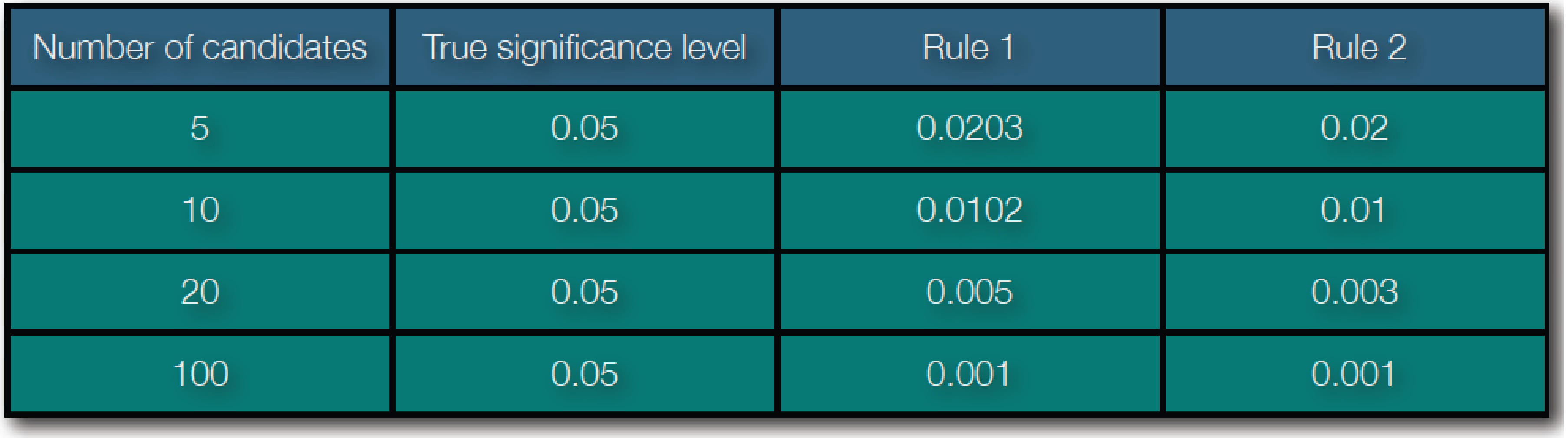
	Stepwise regression	Maximum R²	Max-min[t]
Correct variable selected	82%	75%	36%
Actual Type I Error (when using a 5% critical value)	30%	53%	81%

Data Mining: the bad

- Lovell’s findings:
  - All three model-selection procedures he tested did very poorly
    - Often found the wrong model
    - Actual sizes of nominal tests were very wrong (over optimistic – too often rejected a true null of zero)
    - The max-min t test approach is particularly misleading
  - His key suggestion to deal with this:
    - Due to sampling variation, some significant t-stats/High R-squared values are inevitable, some spuriously
      - Naïve data mining is extremely sensitive to spurious results
    - To avoid spurious selection, critical values (or significance levels) should be made stricter for the degree of search
      - To achieve a true 1% probability of type one error, one should use a much smaller nominal significance level (e.g. 0.01%) depending on how many times the data was reused in the specification search

Suggestions:

- Admit, describe iterations
- Adjust test sizes for iterations
- Lovell suggested two rules:



Number of candidates	True significance level	Rule 1	Rule 2
5	0.05	0.0203	0.02
10	0.05	0.0102	0.01
20	0.05	0.005	0.003
100	0.05	0.001	0.001

- Hoover and Perez (2008) do not agree that this is the best approach. More below

Data Mining: the inevitable

- The seminal contribution is from a Bayesian, Leamer (1978, 1983)
- Main ideas (very roughly)
  - Presenting just the best model is misleading/unconvincing since every econometrician knows all other econometricians (except themselves, of course…) use unwarranted data-mining
  - Thus: present all possible alternative specifications
  - The only validly “significant” results are those that are robust against any possible alternative.
    - In practice: report the min and max of every coefficient in every sensible/feasible possible estimation variant. This is called “extreme bounds analysis”

Data Mining: the inevitable

- Main Criticism of Extreme Bounds Analysis by Hoover and Perez (2000))
  - Again, due to sampling variation, there is no guarantee that the true DGP will be robust to all possible variants.
  - Expecting robustness against *well-specified* alternatives models is reasonable, but not against mis-specified alternatives:
    - When a model is mis-specified, all coefficients estimators are inconsistent and all test statistics are invalid, thus finding an estimate of a “true variable” that is insignificant or “the wrong sign” in a mis-specified model cannot necessarily be taken as evidence against it.
    - This mirrors the discussion we had with regards to unit root testing: even if a process is random-walk without drift, fitting a high enough polynomial in time (in this case, a mis-specification) will result in the rejection of difference stationarity.
  - Thus, while the Leamer approach might never give false positives (type 1 errors), it would almost certainly give many false negatives (type 2 errors)

Data Mining: the warranted (Spanos 2000)

- Basic assumption: There *is* a true DGP although we can never know when we have uncovered it
  - "A good specification-search methodology is one in which the truth is likely to emerge as the search continues on more and more data"
  - Data-mining in some form is essential to uncover the DGP - "The only issue is whether any particular data mining scheme is a good one"
- The *general-to-specific* approach (also called the LSE approach) to data mining argues that it is a necessary and useful approach to uncovering the true DGP
  - I have loosely referred to this approach in most of our empirical analyses
  - The central idea is that, when it is possible, to start with a *General Unrestricted Model (GUM)* of the DGP
    - A specification so flexible and general that the true DGP must be contained in the GUM
  - Given a GUM, all statistical tests are valid, and the GUM can be reduced in a philosophically and statistically defensible way to a parsimonious restricted model that encompasses all other models, which then represents the best estimate of the (local) DGP

Data Mining: the warranted

- The *general-to-specific* approach is top-down proposed as an alternative to the bottom-up approach of Lovell (1983)
    - He started with small models and searched over them. I.e. most of the models were non-nested, so could not be compared "on equal footing"
- Why should the general-to-specific approach work better?
    - The idea of the general-to-specific approach is to start with a large (almost certainly over-specified) model (the GUM) that nests all possible smaller models
    - Reducing from such a GUM is likely to work better, and several simulation studies have showed evidence that it does
    - The reason why it should work better is well articulated by Spanos (2000), by distinguishing between
        - Mis-specification tests, and
        - Primary/economic hypothesis tests

Mis-specification tests vs hypothesis tests

- The main purpose of econometrics is to provide evidence for or against *economic* hypotheses. E.g.
  - Does PPP hold?
  - How large are the returns to education?
  - Is there evidence of collusion between firms?
- In any estimated model, such empirical tests are only valid (unbiased/consistent) if the statistical model of the data is adequate
- We test for the adequacy of a model (*before* we attempt to test for economic hypotheses) by using *mis-specification tests.*
- How are they different?
  - An economic hypothesis tests asks whether there is evidence for or against certain claim in economic theory
  - A mis-specification tests asks whether the empirical model is statistically congruent with the data, *whatever* its economic content

The general-to-specific approach

- Spanos (2000), Hoover and Perez (2000) and Hendry and co-authors all subscribe to this approach as a probabilistic reduction from a GUM to a parsimonious, encompassing empirical model on which economic hypotheses can be validly tested
- The GUM is typically very over specified (poorly estimated coefficients), and hence cannot give sharp answers to questions of economic interest
  - This means reduction of the GUM to a more precise, smaller model is required
  - This is done by successive reductions of the model using mis-specification tests as pruning devices
  - As there may be different paths to reducing a model from GUM to final parsimonious form, this is not an obvious approach
  - Automated versions of the approach trace a large number of paths, which usually leads to many competing candidate final models
  - The candidates are then evaluated relative to each other via *encompassing tests*

The general-to-specific approach

- This literature distinguishes between two very different concepts:
  - The *statistical model* of the data
  - The *economic model* that is tested against the data
- For any test of an economic hypothesis on an estimated model to be valid, the estimated statistical model must be congruent with the data
- The first step for an empirical researcher is therefore to construct a statistical model of the data generating process that is unrelated to the economic model of interest
- In this school of thought, a model is considered (tentatively) to be adequate if it is *congruent* with the data in the following sense:
  1) It yields results consistent with the measuring system (e.g. no negative unemployment rates)
  2) Residuals are white noise and true innovations w.r.t. the measured variables
  3) Coefficient estimates are stable w.r.t. the sample selection

The general-to-specific approach

- A estimated model that does not have these basic characteristics is obviously mis-specified, and cannot inform on any economic hypothesis (as all the coefficients are unreliable estimates of truth)
- An estimated model that does have these characteristics might be valid, but still may be too uncertain (too over-specified, so that coefficient estimates are too imprecise) to give strong evidence on economic hypotheses
- So the goal becomes to find the smallest model that is

1) Congruent with the data

2) Encompasses all other models – i.e. can explain the results of all other models

The general-to-specific approach

- The ideal is obviously infeasible:
    - Estimate a model with all possible variables, interactions, lags, non-linearities
    - This must contain the true model, so we should be able to find it from there
- The feasible strategy:
    - Estimate the largest feasible model
    - Test that it is congruent with the data with a battery of specification tests for the statistical model:
        - White noise residuals
        - Appropriately exogenous regressors
        - Stable estimated coefficients
            - Passing these tests does not prove congruency
            - Failing these tests rejects congruency – you need to start with an even larger model

The general-to-specific approach

- The feasible strategy:
  - If the large model is congruent, start reducing
    - Restrict one of the least significant coefficients to zero (Darwinian approach: kill the least significant coefficient)
    - Re-evaluate congruency (does the restriction lead to a loss of information?)
    - If congruency fails, reject the restriction and move onto the next least significant
    - Continue reducing the model until congruency fails
  - This protocol will lead to a set of smallest models that remains congruent
    - There may be many different paths – sampling variation means the “least significant” coefficient is not well identified
    - The most universal general-to-specific approach traces many paths of reduction (in terms of where one starts)
  - Test for encompassing of all terminal models

A test of the general-to-specific approach

Krolzig, H.M., 2003. General-to-specific model selection procedures for structural vector autoregressions. *Oxford Bulletin of Economics and Statistics*, *65*, pp.769-801.

- Monte Carlo exercise on I(0) SVAR
  - Construct SVAR/VAR from universe of 3 variables
  - *PcGets* is an automated implementation of the approach (only in OxMetrics last time I checked)

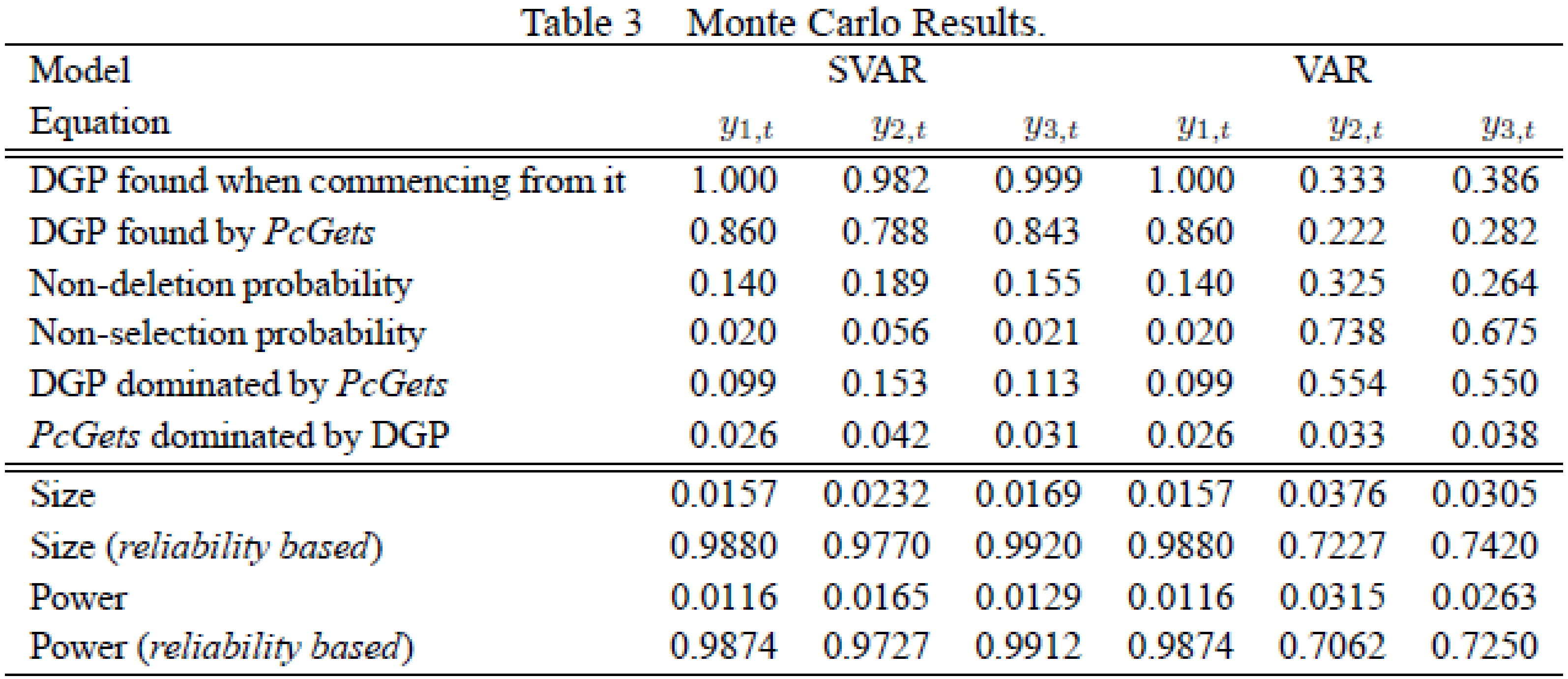


Table 3 Monte Carlo Results.

Model	SVAR			VAR		
Equation	$y_{1,t}$	$y_{2,t}$	$y_{3,t}$	$y_{1,t}$	$y_{2,t}$	$y_{3,t}$
DGP found when commencing from it	1.000	0.982	0.999	1.000	0.333	0.386
DGP found by *PcGets*	0.860	0.788	0.843	0.860	0.222	0.282
Non-deletion probability	0.140	0.189	0.155	0.140	0.325	0.264
Non-selection probability	0.020	0.056	0.021	0.020	0.738	0.675
DGP dominated by *PcGets*	0.099	0.153	0.113	0.099	0.554	0.550
PcGets dominated by DGP	0.026	0.042	0.031	0.026	0.033	0.038
Size	0.0157	0.0232	0.0169	0.0157	0.0376	0.0305
Size (*reliability based*)	0.9880	0.9770	0.9920	0.9880	0.7227	0.7420
Power	0.0116	0.0165	0.0129	0.0116	0.0315	0.0263
Power (*reliability based*)	0.9874	0.9727	0.9912	0.9874	0.7062	0.7250

Counter argument:

- Note that the analytic and Monte Carlo results of the general-to-specific approach are explicitly based on I(0) systems, although it is claimed that the method should give sensible answers in CI(1,1) systems as well
- Lutkepohl (2007) argues that, especially in co-integration studies, general-to-specific modelling is not the procedure that is followed “globally” in a study, although it may be done on components of the modelling

Lütkepohl, H., 2007. General-to-specific or specific-to-general modelling? An opinion on current econometric terminology. *Journal of Econometrics*, *136*(1), pp.319-324.

Lutkepohl (2007):

- His description of general practice mirrors how we did co-integration analysis:
  - Step 1: test for unit roots individually
    - Recently, unit root tests can be done jointly
  - Step 2: among I(1) variables, investigate co-integrating relationships
    - He suggests that theoretically and practically, small subsystems should be considered first:
      - Bad small sample properties of large systems
      - “a better chance of finding interesting relationships”
  - Step 3: a small VAR/VEC model is used, reduced by the general-to-specific approach procedure
    - If the initial large model (as before) is not congruent with the data, even general-to-specific approach needs to augment
    - Spanos (2000), however, argues that adding lags is fundamentally different than adding variables to a system. Increasing the lag structure should not violate any of the rules
  - Step 4: extensions of the model are considered to check for robustness

My opinion

- Lutkepohl is right that general-to-specific approach is not a good description of what econometricians actually do
- But the general-to-specific idea is not about how things are *typically* done but a suggestion of how they *should* be done
- The poor estimation of a large model is indeed a concern, and we have seen this in our explorations
- However, he provides only concepts and ideas on what is *often done* without providing evidence that it works as well/better than a general-to-specific approach
- Whether the general-to-specific approach works in a cointegration setting would have to be tested in a Monte Carlo “horse race” against other alternatives

A Problem with automated general-to-specific modelling

- Last time I checked, it was only implemented by the originators in OxMetrics by Jurgen Doornik
  - OxMetrics is a proprietary econometrics program developed at Oxford
  - See http://www.doornik.com/

- I like the agnostic, systematic approach, even if “done by hand” in smaller models than they suggest.

Model Selection

- Even in the ideal setting, the general-to-specific reduction approach can lead to a variety of different final models
  - Depending on the sequence of reductions, different paths may lead to different final models
- How do we choose between them?
- The approach suggested by authors in the general-to-specific camp is to test for whether one (or more) models *encompass* the others
  - If model 1 encompasses model 2, it explains all the findings of model 2
- If one model encompasses all the others, it is the preferred model to use to test economic hypotheses on.

Encompassing

- Model selection is very complicated, and deeply depends on the situation
  - In the multiple equation VAR setting with an obvious set of variables, it is usually relatively simple – lag length is the only issue
  - When the choice of explanatory variables is at issue, it is much more complicated
- Let’s construct a simple example to illustrate the idea

Encompassing

- Suppose we used a general-to-specific approach to reduce the statistical model of the DGP to two congruent models:

$$M_1: y_t = \beta_0^{M_1} + \beta_x^{M_1} x_t + e_t^{M_1}$$

$$M_2: y_t = \beta_0^{M_2} + \beta_z^{M_2} z_t + e_t^{M_2}$$

- These models are not nested – we can't test one as a restriction of the other
- An encompassing test would be to form the union of the two models:

$$M_3: y_t = \beta_0^{M_3} + \beta_x^{M_3} x_t + \beta_z^{M_3} z_t + e_t^{M_3}$$

- Provided M_3 is also congruent, we can test if there is evidence of encompassing:
  - If the null hypothesis that $\beta_z^{M_3} = 0$ cannot be rejected but the null hypothesis that $\beta_x^{M_3} = 0$ can be rejected, then model M_1 is a valid reduction of model M_3 but model M_2 is not. In other words, model M_1 encompasses model M_2 - it is congruent and encompasses (explains) the results of model M_2: z_t is only significant in model M_2 because the "true" variable x_t was erroneously omitted.
  - If neither zero null hypothesis in M_3 can be rejected, then neither of the smaller models encompasses the other, and the union of the two models is the appropriate model

Summary

- Doing econometrics well is difficult
    - So much endogeneity and simultaneity, so many options, so little data
    - Somewhere between an art and a science
    - We should strive to be as scientific in our production and reporting of our artworks as possible
- One must take the risks very seriously
    - In one's reading of other's work as well as in one's own
    - Read widely and critically on this topic and form and continuously update your own best practice
        - None of the econometric high priests are right about everything, none of them are wrong about everything
        - Kennedy (2002) provides an amusing set of ten commandments for good data work that speak to a lot of issues. He gives very useful advice and pulls together a large variety of opinions and work on the issues
- Some data-mining is inevitable
    - Do it systematically, carefully, honourably
    - Honestly report as much of the process as is feasible/important
    - Use solid statistical and economic reasoning to argue for your "best model" but give the readers the truth about its robustness/stability
    - Our job is (or should be) to agnostically test economic theories without preconceptions and our statistical practices should reflect this as best as possible