The probability of finding the Data Generation Process after a t^2-testing sequential procedure

J. Campos
Economics Department, University of Salamanca, Spain.
(e-mail: jcampos@usal.es)

January 28, 2004

Abstract

Recently developed algorithms for model selection proceed by ordering the variables by their absolute $t-$values. So much testing involved in choosing the final model, and the generality of the initial model have been questioned in the literature. This paper shows that increasing model generality and repeated t^2-testing do not reduce the chances of finding the DGP. However, if somewhere along the procedure an irrelevant variable is incorrectly included then all following variables in the sequence are retained with probability one.

JEL classification: C12, C22, C32, C51, C52.
Keywords: General-to-Specific. Data mining. LSE econometrics. Model selection.

I am grateful to D.F. Hendry for helpful comments. The derivations in this paper were motivated by reading his joint work with H.-M. Krolzig. M. Knott found an infelicity in an earlier version of this paper, which led to particularize, the nonetheless more general and correct results provided in that version, to be applicable to the questions I had intended to address. All errors are my own responsibility. I acknowledge use of GiveWin 2.10 and Mathematica 4.2.

1 Introduction

Recently developed automated procedures for model selection (see Hoover and Perez (1999) and Hendry and Krolzig (1999)) start by ordering the variables in the initial model by their absolute $t-$values. We would like to examine the consequences of such action on the probabilities of retaining in and deleting from the initial model relevant and irrelevant variables. Automated procedures are more sophisticated than the context we are about to present. In these procedures variables are ordered but the decision on whether they are finally included or not is made upon the results obtained from applying diagnostic tests. An important feature of both HP and HK is that variable selection is accomplished by means of individual and block hypothesis testing. Lovell (1983) found in a Monte Carlo study that $t-$testing is not a useful procedure for detecting relevant variables. Hansen (1999) questions the consistency of model selection procedures based on hypothesis testing, arguing that those procedures lead to overparameterized models even in large samples. He proposes to use consistent information criteria to choose between all possible models which can be formulated from a set of potential variables. However, to reduce dimensionality he recommends to estimate the most general model, order its variables by their absolute $t-$values and delete the insignificant variables one at a time till a model with about 10 regressors is obtained. He suggests to formulate all possible models (2^{10}) from these 10 regressors and choose that model with smallest Schwarz information criterion. Campos et al. (2003) derive and compute the probability of finding the DGP when all variables are irrelevant in Hansen's procedure. It is found that this probability increases with pre-selection and with the penalty parameter in the Schwarz information criterion.

What we would like to examine below is whether ordering the variables according to their t^2-values alters Lovell (1983) result, and to what extent the claimed pernicious effect of sequential testing applies. The results below, which are derived assuming that the tests-statistics are mutually independent, and for the simple cases in which up to three variables in the general model are relevant, question common beliefs on the consequences of repeated testing and model generality. These results are more directly applicable to Hoover and Perez (1999) and Hansen (1999) because of their intensive use of $t-$testing. The derivations appear to generalize to models with more than three relevant variables. However, relaxing the independence assumption complicates the derivations.

Using the theory on order statistics we derive particular expressions for the following probabilities, given that the set of variables with lower t^2-values have already been excluded. First, when all potential variables are irrelevant, Section 2 finds the probabilities of correct deletion and of incorrectly retaining those variables with largest t^2-values. Second, Section 3 finds the probabilities of retaining the most significant variables when the initial model includes up to three relevant variables. Because t^2-statistics associated to relevant variables have non-central distributions their means are larger than the means of those statistics corresponding to irrelevant variables. Hence, we expect the relevant variables to be also the most significant, and so hope that the derived probabilities of the most significant variables coincide with the probabilities of retaining the relevant variables. If that is so, contrary to common knowledge, it is found that the probabilities of retaining relevant variables are not lower for more general models, and that repeated single t^2-testing does not reduce the probability of selecting the relevant variables. Some conventional results also hold in this context: the probability of correct inclusion increases with the number of relevant variables and with departures from the null hypothesis. The results are first derived assuming that all non-central t^2-statistics have the same distribution which is not the case in practice. However, inspection of the particular case of two relevant variables suggests that the same results hold when the t^2-statistics have different distributions.

2 Models with no relevant variables

Automated procedures for model selection permit increasing model generality. In this Section we would like to address two issues related to the generality of the initial model: (i) whether considering larger sets of potential variables affects size and power of tests leading to incorrect exclusion and inclusion, and (ii) whether sequentially eliminating potential explanatory variables reduces the probability of finding the DGP. The answer to (i) is provided by deriving the probabilities of correctly excluding and of incorrectly including the set of variables with largest t^2-values, given that the remaining variables have already been excluded, in a sequential testing procedure when all variables are irrelevant. (ii) is examined by comparing sequential conditional probabilities.

Let us consider the following general initial model with k regressors:

$$y_t = \sum_{i=1}^{k} \gamma_i x_{ti} + \varepsilon_t. \tag{1}$$

Let us suppose that we wish to discover what regressors are relevant by using a hypothesis testing sequential procedure in which regressors have been arranged by their squared $t-$values in an increasing order of magnitude (i.e., the first variable in the sequence is that with smallest t^2-value and the last variable is that with largest t^2-value). What we are going to do is to derive the probabilities of excluding and retaining variables when all regressors are irrelevant, and all t^2-statistics are mutually independent. Let us denote the ordered mutually independent t^2-statistics computed from T observations by $0 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k < \infty$. Hence, the joint density of those τs is:

$$D_{\tau_1,\ldots,\tau_k}(\tau_1,\ldots,\tau_k) = \begin{cases} k! \prod_{i=1}^{k} f(\tau_i) & 0 \leq \tau_1 \leq \cdots \leq \tau_k \leq \infty \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $f(\tau_i)$ is the marginal density of the ith $\tau-$statistic (see e.g., Hogg and Craig (1970) and David (1981)).

Denoting by $p = F\left(\xi_p\right)$, where $F(\cdot)$ is the cumulative distribution function (cdf) of a central distribution with density $f(\cdot)$, and by:

$$P_r = P\left[\tau_i \leq \xi_p, i = 1,\ldots,r\right] = P\left[\tau_r \leq \xi_p\right]$$

integration of (2) yields $P_k = p^k$ and:

$$P_{k_1} = \sum_{i=0}^{k-k_1} \frac{k!}{i!\,(k-i)!} p^{k-i}\theta^i := S_{k,k_1} \tag{3}$$

where $\theta = 1 - p$.

We wish to derive first the probabilities of correct exclusion. The probability of excluding the last $k - k_1$ variables given that the remaining k_1 have been excluded is:

$$P\left[\tau_i \leq \xi_p, i = k_1 + 1,\ldots,k \mid \tau_{k_1} \leq \xi_p\right] = \frac{P_k}{P_{k_1}} = \frac{p^k}{S_{k,k_1}} \tag{4}$$

Variables	Probabilities of correct exclusion			
	$k = 40$		$k = 60$	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
All	0.128	0.669	0.046	0.547
Three most significant	0.149	0.669	0.071	0.549
Two most significant	0.190	0.674	0.110	0.560
Most significant only	0.322	0.712	0.241	0.623

Table 1: Conditional probabilities of correctly excluding all, the last three, the last two, and the kth variable only.

which is a function of the probability p to the left of the critical value, of the total number k of variables in the initial model, and of the number k_1 of irrelevant variables already excluded. For fixed k, the sequence of probabilities in (4) defined for $k_1 \in [1, k-1]$ can be interpreted as follows: $k_1 = 1$ yields the probability of excluding the $k-1$ variables with highest t^2-values, once we have tested for the significance of the first variable (the variable with smallest t^2-value) and we have excluded it; $k_1 = 2$ yields the probability of excluding the $k-2$ variables with highest t^2-values, once we have tested for the individual significance of the first two variables (those variables with smallest t^2-values) and we have excluded them; and so on.

Because (3) implies that $P\left[\tau_1 \leq \xi_p\right] = S_{k,1} = 1-(1-p)^k \neq 0$, for all $p > 0$, sequential testing does not reduce the probability of correct exclusion since for $j = 2, \ldots, k$:

$$\begin{aligned} P\left[\tau_i \leq \xi_p, i = 1, \ldots, k\right] &< P\left[\tau_i \leq \xi_p, i = j, \ldots, k \mid \tau_{j-1} \leq \xi_p\right] \\ &< P\left[\tau_i \leq \xi_p, i = j+1, \ldots, k \mid \tau_j \leq \xi_p\right]. \end{aligned}$$

In addition, probabilities of correct exclusion are larger when testing at lower significance levels, and seem to be smaller in more general models. Table 1 provides probabilities, computed as in (4), of correctly excluding variables in this sequential procedure for models with 40 and 60 variables, and when testing has been carried out at the 5% and 1% significance levels. The probability of excluding the variables with highest t^2-values increases with the number of variables with lower t^2-values previously excluded (k fixed but k_1 increasing). For instance, for $\alpha = 0.01$ the probability of excluding all $k = 40$ variables is 0.669, and the probability of excluding the most significant variable when all remaining variables have already been excluded is

0.712. So, from Table 1 we find that probabilities of correct exclusion: (i) are not reduced by sequential testing; (ii) decrease with model generality; and (iii) are larger at lower significance levels, in agreement with the results found in the literature (see e.g., Hoover and Perez (1999) for their model 1).

(3) also allows us to get insight on overparameterization after a sequential testing procedure. The probability of incorrectly including the remaining variables given that the first k_1 have already been correctly excluded is:

$$\begin{aligned} \mathsf{P}\left[\tau_i \geq \xi_p, i = k_1 + 1, \ldots, k \mid \tau_{k_1} \leq \xi_p\right] &= \frac{\mathsf{P}\left[\tau_{k_1} \leq \xi_p \leq \tau_{k_1+1}\right]}{\mathsf{P}\left[\tau_{k_1} \leq \xi_p\right]} \\ &= \frac{k! p^{k_1} \theta^{k-k_1}}{(k-k_1)! k_1! S_{k,k_1}}. \end{aligned} \tag{5}$$

(5) implies that sequential testing leads to larger probabilities of incorrect inclusion because $\mathsf{P}\left[\tau_i \geq \xi_p, i = k_1 + 2, \ldots, k \mid \tau_{k_1+1} \leq \xi_p\right]$ is greater than $\mathsf{P}\left[\tau_i \geq \xi_p, i = k_1 + 1, \ldots, k \mid \tau_{k_1} \leq \xi_p\right]$. Furthermore, once a variable has been included the following variables in the sequence are also included with probability one because $\mathsf{P}\left[\tau_i \geq \xi_p, i = k_1 + 2, \ldots, k \mid \tau_{k_1} \leq \xi_p, \tau_{k_1+1} \geq \xi_p\right] = 1$. Hence, sequential testing establishes a trade-off between correct exclusion and incorrect inclusion. Table 2 shows computed probabilities of incorrect inclusion. By comparing Tables 1 and 2 we notice that probabilities of correct exclusion are larger than the probabilities of incorrect inclusion when testing at the 1% significance level. When $k = 40, 60$ for $\alpha = 0.01$ those discrepancies are as large as 0.67 when inclusion-exclusion of the last $k - 1$ variables is considered for $k = 40$, and reduced to 0.42 when inclusion-exclusion of the last variable is judged. These numbers drop to 0.55 and 0.24, respectively, when $k = 60$ but those differences are not negligible. In general, (4) and (5) imply that the discrepancy between the probabilities of correctly excluding and of incorrectly including the last $k - k_1$ variables is positive for models with even 80 variables when testing proceeds at significance levels not greater than 1%. So, the benefit of sequential testing can offset the evil of incorrect inclusion.

In addition, (5) converges to zero as p approaches unity but increases when both k and k_1 become larger. The numbers in Table 2 illustrate that the discrepancy between the probabilities of keeping the most significant but irrelevant variables increases with model generality, that we do much better by testing at lower significance levels, and that it is less likely to retain more than fewer irrelevant variables. By testing at the 1% significance level we

Variables	Probabilities of incorrect inclusion			
	k = 40		k = 60	
	α = 0.05	α = 0.01	α = 0.05	α = 0.01
Five most significant	0.035	0.000	0.110	0.000
Four most significant	0.095	0.000	0.210	0.003
Three most significant	0.215	0.007	0.355	0.019
Two most significant	0.410	0.054	0.541	0.101
Most significant only	0.678	0.288	0.759	0.377

Table 2: Conditional probabilities of incorrectly including up to the last five variables.

would expect

to include two variables about 5% of the time. Diagnostic testing may explain why Hoover and Perez (1999) obtain almost no falsely included variables at the 1% level (see also Hendry and Krolzig (1999)). Table 2 also indicates that the probability of incorrectly including the last five variables when all remaining variables have already been excluded is negligible, but it is as high as 29% when testing for the significance of the last variable at the 1% significance level for k = 40. Notice that the choice of significance level is crucial for keeping these probabilities low.

In this Section we have derived formulae for computing conditional probabilities of correct exclusion and inclusion when all variables in the initial model are irrelevant. We have found that: (i) the probability of correct exclusion is not reduced by sequential testing, it is larger when testing at lower significance levels, and it is smaller for more general initial models; (ii) the probability of incorrect inclusion is enhanced by sequential testing, it is sensibly reduced when testing at smaller significance levels, and increases with model generality; (iii) the benefit from sequential testing more than offsets the evil of incorrect inclusion; and (iv) once a variable has been incorrectly included all variables following that in the sequence are included with probability one. The next Section deals with models which include relevant as well as irrelevant variables. Curiously, it is found that more general models do not lead to lower probabilities of keeping relevant variables. However, as we may expect, the probability of choosing relevant variables is likely to be enhanced by increasing the number of relevant variables in the initial model.

3 Models with relevant variables

In what follows model (1) is assumed to include regressors which are relevant for explaining variable y. Regressors have been arranged according to their t^2-values in an increasing order of magnitude. We assume that the t^2-statistics are mutually independent. Subsection 3.1 derives the probabilities of retaining relevant variables, given that all irrelevant variables and less significant but relevant variables have already been excluded. It is assumed that the t^2-statistics associated to the relevant variables are identically distributed, and that testing proceeds using the same critical value for all hypotheses in the sequence. Subsection 3.2 relaxes the assumption of identical distribution, but, except for a simple example, it keeps the assumption that all test-statistics have distributions with the same degrees of freedom. Degrees of freedom determine the critical values in a testing sequence. Critical values are nearly constant when testing for the significance of irrelevant variables but they may change substantially when the model includes variables which are relevant. However, the example provided below seems to indicate that probabilities of correct inclusion, computed using the same critical values for all subsequent hypotheses, provide a lower bound to the same probabilities calculated using the correct critical values.

3.1 Probabilities when t^2-statistics are identically distributed

Let us sequentially consider model (1) with one, two and three relevant variables. Let us denote by $\mathsf{f}(\cdot)$ and $\mathsf{g}(\cdot)$ the densities of a central and a non-central distribution, respectively. The probability of retaining variable s given that all previous variables in the sequence have been excluded, in a model with r relevant variables, is:

$$\mathsf{P}_r\left\{\tau_s \geq \xi_p \mid \tau_{s-1} \leq \xi_p\right\} = 1 - \frac{\mathsf{P}_s}{\mathsf{P}_{s-1}}. \tag{6}$$

We next derive those probabilities for each model.

3.1.1 Models with one relevant variable

We consider model (1) with one relevant variable. The density of the ordered t^2-statistics is:

$$D_{\tau_1,\ldots,\tau_k}(\tau_1,\ldots,\tau_k) = (k-1)!\prod_{j=1} g(\tau_j)\prod_{i\neq j} f(\tau_i)$$

for $0 \leq \tau_1 \leq \cdots \leq \tau_k < \infty$, and zero otherwise. Integration yields $P_k = p^{k-1}q$, $P_{k-1} = p\lambda + q\varphi_{1,1}\, p^{k-2}$, and:

$$P_{k-2} = p^{k-3}\left[p\{1+(k-2)\theta\} + \frac{1}{2}(k-1)(k-2)q\theta^2\right]$$

where $\lambda = 1-q$, $\varphi_{r,s} = rp + (k-s)\theta$, and $q = G\left(\xi_p\right)$; with $G(\cdot)$ being the cdf of a non-central distribution with density $g(\cdot)$. So:

$$P_1\left[\tau_k \geq \xi_p \mid \tau_{k-1} \leq \xi_p\right] = 1 - \frac{pq}{p\lambda + q\varphi_{1,1}} \tag{7}$$

for $k = 2,3,\ldots$. Strictly speaking (7) is the probability that the most significant variable is correctly retained when less significant variables have already been excluded. We would expect τ_k to correspond to the t^2-value of the relevant variable, in which case (7) would be the probability of retaining the relevant variable provided the irrelevant variables have already been excluded. As an example, notice that non-central χ^2 and F distributions have larger means than their central counterparts, and hence we would expect larger t^2-values for variables which are more relevant.

(7) increases as k becomes larger and hence the probability of retaining the relevant variable is not smaller for models with more variables (see Table 3). However, once a variable has been included the variable with the next highest t^2-value is also included with probability one, no matter whether that variable is relevant or not. In addition, model generality leads to larger probabilities of incorrect inclusion as the following expression indicates: The probability of including the most significant irrelevant variable is:

$$P_1\left[\tau_{k-1} \geq \xi_p \mid \tau_{k-2} \leq \xi_p\right] = 1 - \frac{\left[p\lambda + q\varphi_{1,1}\right]p}{p\{1+(k-2)\theta\} + \frac{1}{2}(k-1)(k-2)q\theta^2} \tag{8}$$

which increases with k.

3.1.2 Models with two relevant variables

We next increase the number of relevant variables in model (1) to two. The density of the τs is:

$$D_{\tau_1,\ldots,\tau_k}(\tau_1,\ldots,\tau_k) = 2(k-2)! \prod_{j<h} g(\tau_j)\, g(\tau_h) \prod_{i\neq j\neq h} f(\tau_i)$$

for $0 \leq \tau_1 \leq \cdots \leq \tau_k < \infty$, and zero otherwise. By integrating the latter density we obtain $P_k = p^{k-2}q^2$, $P_{k-1} = (2p\lambda + q\varphi_{1,2})\, p^{k-3}q$, and $P_{k-2} = (p^2\lambda^2 + q^2 + 2pq\lambda\varphi_{1,2} + \frac{1}{2}(k-4)\, q^2\theta\varphi_{2,1})\, p^{k-4}$ for $k = 3, 4, \ldots$. So, $P_2(\tau_k > \xi_p \mid \tau_{k-1} \leq \xi_p)$ is greater than $P_1(\tau_k > \xi_p \mid \tau_{k-1} \leq \xi_p)$ which implies that the probability of correctly including the most relevant variable increases with the number of relevant variables in the initial model. In addition, $P_2(\tau_k > \xi_p \mid \tau_{k-1} \leq \xi_p)$ is greater than $P_2(\tau_{k-1} > \xi_p \mid \tau_{k-2} \leq \xi_p)$ and hence sequential testing does not lead to lower probabilities of choosing the relevant variables as we proceed along the sequence of tests.

3.1.3 Models with three relevant variables

In what follows we consider a situation in which we have three relevant variables in model (1) and look at the probabilities of retaining those variables. For $0 \leq \tau_1 \leq \cdots \leq \tau_k < \infty$, the density is:

$$D_{\tau_1,\ldots,\tau_k}(\tau_1,\ldots,\tau_k) = 3!\,(k-3)! \prod_{j<h<m} g(\tau_j)\, g(\tau_h)\, g(\tau_m) \prod_{i\neq j\neq h\neq m} f(\tau_i)$$

and zero otherwise. So, $P_k = p^{k-3}q^3$, $P_{k-1} = (3p\lambda + q\varphi_{1,3})\, p^{k-4}q^2$, $P_{k-2} = (3p^2\lambda^2 + pq^2 + 3pq\lambda\varphi_{1,3} + \frac{1}{2}(k-4)\, q^2\theta\varphi_{2,3})\, p^{k-5}q$, and:

$$P_{k-3} = \Big(p^3\lambda^3 + q^3 + 3p^2q\lambda^2\varphi_{1,3} + 3pq^2\lambda\Big(1 + \frac{1}{2}(k-5)\,\theta\varphi_{2,2}\Big) + (k-6)\Big(1 + \frac{1}{2}(k-5)\,\theta + \frac{1}{6}(k-5)(k-4)\,\theta^2\Big)q^3\theta\Big)\, p^{k-6}.$$

for $k = 4, 5, \ldots$. We obtain the same results: sequential testing does not reduce the chance of retaining the relevant variables, and more relevant variables enhance the probability of being selected.

Summarizing, in this Section we have found that: (i) probabilities of retaining relevant variables increase with the total number of variables in the initial model; (ii) sequential testing does not reduce the probability of correctly including the next relevant variable, even if the previous relevant variable has been incorrectly excluded; (iii) probabilities of correct inclusion are larger when the number of relevant variables in the model is larger; and (iv) probabilities of incorrect inclusion increase with the total number of variables in the initial model.

3.1.4 Some calculations

In what follows we wish to compare those probabilities in this Section for models with $k = 40$ and $k = 60$ variables, which have been estimated using $T = 140$ observations. We wish to calculate the probability of finding variables with t^2-values greater than the critical value corresponding to testing at the 1% significance level, when variables are relevant but their t^2-ratios have a distribution close to the null, and to compare them with the probabilities of retaining those variables when that distribution is farther away from the null. The values of p and q are associated to the central and non-central χ^2 distributions, respectively. $p = 0.99$. To obtain q we consider an approximation to the non-central distribution of the t^2-statistic around zero. Let us assume that variable x_i is relevant, denote its parameter by γ_i and by σ_i^2 its variance. In addition, let σ_ε^2 be the variance of the disturbances in the model. Independence of regressors and appropriate additional conditions imply that t_0^2 is approximately distributed as a non-central χ^2 with 1 degree of freedom and non-centrality parameter $\psi = T\sigma_\varepsilon^{-2}\gamma_i^2\sigma_i^2$. To compute q we approximate that $\chi^2(1;\psi)$ by the variable $(1+2\psi)^{-1}(1+\psi)\chi^2$, which is a central χ^2 with $\nu = (1+2\psi)^{-1}(1+\psi)^2$ degrees of freedom. So:

$$q = P\left[t_0^2 \leq \xi_p\right] \simeq P\left[\chi^2(1;\psi) \leq \xi_p\right] \simeq P\left[\frac{1+\psi}{1+2\psi}\chi^2(\nu) \leq \xi_p\right] \qquad (9)$$

and we expect a relevant variable to have a t_0^2-value around $(1+2\psi)^{-2}(1+\psi)^3$.

When the null is true and so a variable is irrelevant, the critical value for testing at the 1% significance level is 6.63. So, we consider values of ψ which lead to the two situations we chose to examine: (i) variables with t_0^2-values close to 6.63 and hence with associated parameters close to the null, and (ii) variables with t_0^2-values farther away from 6.63 and hence with associated parameters farther away from the null. The first value of ψ is found by

Non-centrality parameters and Variables	Probabilities of retention					
	1 relevant		2 relevant		3 relevant	
	k = 40	k = 60	k = 40	k = 60	k = 40	k = 60
ψ = 24.49:						
Most significant	0.54	0.58	0.67	0.69	0.74	0.75
Penultimate significant			0.31	0.36	0.44	0.47
Antepenultimate significant					0.18	0.22
Irrelevant most significant	0.15	0.22				
ψ = 29.50:						
Most significant	0.65	0.68	0.77	0.78	0.83	0.84
Penultimate significant			0.44	0.48	0.59	0.61
Antepenultimate significant					0.30	0.34
Irrelevant most significant	0.19	0.26				

Table 3: Probabilities of retaining variables at the 1% significance level.

solving for ψ the cubic $(1 + 2\psi)^{-2}(1 + \psi)^3 = 6.63$ which yields $\psi = 24.49$. We obtain a second value of $\psi = 29.50$ by solving the same cubic but equated to 7.88 so that in (ii) we expect variables to show a t^2-value of about 7.88 in which case the null would be rejected.

q is approximately 0.5522 and 0.4008 for $\psi = 24.49$ and 29.50, respectively. Table 3 shows, for $k = 40$ and $k = 60$, the probabilities of retention when there is only one relevant variable and when there are two and three relevant variables in the potential set, as computed from (6) with P_s, P_{s-1} and P_{s-2} replaced by their appropriate expressions, given that all previous variables in the sequence have already been excluded, testing is carried out at a 1% significance level, and t^2- statistics associated to the relevant variables have the same distribution. In particular, Table 3 shows the probabilities of keeping only the most significant variable, the penultimate significant, and the antepenultimate significant variable. Because all relevant variables have the same distribution we expect all of them to have the same t^2-value. So, for all relevant variables with the same t^2-value their probabilities of being retained are determined by the order in which we consider them. For instance, when there are three relevant variables, the category "Most significant" is to be interpreted as the third variable we consider retaining, given that all irrelevant and the other two relevant variables have been excluded.

A general result is due: probabilities of correct inclusion are not smaller

in more general models. For instance, when there is only one relevant variable in the potential set, the probability of choosing that variable is 0.54 when k = 40 and the first 39 variables have already been deleted, and 0.58 when k = 60 and the first 59 variables have already been deleted. However, other conventional results hold: (i) the chances of choosing relevant variables increase with the total number of relevant variables in the potential set (e.g., 0.54 vs. 0.67 when the potential set includes only one relevant and two relevant, respectively); (ii) the chances of choosing variables with lower t^2-values are smaller (e.g., 0.18 vs. 0.74 which are the probabilities of choosing the least significant only and of keeping the most significant, respectively, when k = 40), or it is more likely to keep the last relevant variable we consider in the testing sequence; and (iii) the probability of retaining relevant variables increases with departures from the null (e.g., 0.74 for ψ = 24.49 vs. 0.83 for ψ = 29.50 which are the probabilities of keeping the most relevant variable when there are three relevant variables in a potential set of 40 variables). Table 3 also records the probability of incorrectly including the most significant irrelevant variable when there is only one relevant variable. Probabilities of incorrect retention are much smaller (e.g. for k = 40 and when the relevant variable is close to being irrelevant, the sequential procedure would include the most significant irrelevant variable 15 out of 100 times). In addition, probabilities of incorrect retention are not smaller when the relevant variables are more important (e.g., 0.15 vs. 0.19 when ψ = 24.49 and ψ = 29.50, respectively).

3.1.5 Compare to Hoover and Perez (1999)

For further illustration we next compare the computed probabilities from (7) with the frequencies of retaining the relevant variable in Hoover and Perez (1999) DGPs 2, 4 and 5. Column 2 in our Table 4 provides the noncentrality parameter associated to each of the relevant variables calculated as $\psi = T\sigma_\varepsilon^{-2}\gamma_i^2\mathbf{b}_i^2$, where T = 140, σ_ε is given in HP's Table 3 as $s.e.r.$, γ_i is the parameter of the ith regressor from HP's Table 3 , and $\mathbf{b}_i$ is taken from HP's Table 2. The remaining columns in our Table 4 show the approximate mean values of $\mathbf{t}_0^2$; the computed probabilities from expression (7) with q as in (9); and the frequencies of retaining relevant variables in HP's Table 7 for DGPs 2, 4 and 5. Testing is carried out at the 1% significance level. The approximate critical value when testing at the 1% significance level is close to 13, and $\mathsf{E}[\mathbf{t}_0^2] > 40$ for the regressors in DGP's 2, 4 and 5, so we would ex-

$\mathcal{DGP}$	ψ	$E[t_0^2]$	Prob.	$\mathcal{HP}$
$y2 = 0.75y2_{-1} + \varepsilon; \sigma_\varepsilon = 85.99$	180	45.5	1	1
$y4 = 1.33x11 + \varepsilon; \sigma_\varepsilon = 9.73$	189	47.6	1	0.999
$y5 = -0.046x3 + \varepsilon; \sigma_\varepsilon = 0.11$	1900	475.6	1	1

Table 4: HP's DGPs, non-centrality parameters, mean values of t-squared, computed probabilities of retention, and HP's frequencies of retention.

pect a large probability of keeping the relevant variable. Column 4 indicates that those variables are kept with probability one, in agreement with HP's frequency of retention recorded in column 5.

3.2 Probabilities when the t^2-statistics have different distributions

We next derive probabilities of retaining relevant variables for models with up to four regressors. t^2-statistics are assumed to be mutually independent, and those associated to the relevant variables have different distributions. However, we assume that the test-statistics have distributions with the same degrees of freedom, so that all subsequent hypotheses are tested using the same critical values. In practice, degrees of freedom increase as we proceed along the sequence, so it would make sense to change the critical values accordingly. What we do next is to consider first a simple model with two regressors, one of which is relevant, allow for the test-statistics in testing for the significance of one variable at a time to have different degrees of freedom, and find the probability of retaining the relevant variable using the correct critical value. That probability is compared to (7), which has been derived using the same critical value for both hypotheses, to illustrate that the probability of retention derived under constant critical values may provide a lower bound to the probability obtained using the appropriate critical values. After that example we consider probabilities with constant critical values.

3.2.1 Two regressors: one relevant variable

Let us denote the critical values associated to testing at the $(1-p)\,\%$ significance level by ξ_{p_1} and ξ_{p_2} for the irrelevant and the relevant variables, respectively, and assume $\xi_{p_2} \leq \xi_{p_1}$. So, $p = F\left(\xi_{p_1}\right) = G\left(\xi_{p_2}\right)$, where $F(\cdot) = G(\cdot)$

are the cdfs of the t^2-statistics associated to the irrelevant and relevant variables, respectively. Let us denote $F\left(\xi_{p_2}\right) = p_2$ and $G\left(\xi_{p_1}\right) = q$. Hence, $p > p_2$, $q > p$, and:

$$P\left(\tau_2 \geq \xi_{p_2} \mid \tau_1 \leq \xi_{p1}\right) = \frac{\int_{\xi_{p_2}}^{\infty} \int_{0}^{\xi_{p_1}} D_{\tau_1,\tau_2}(\tau_1,\tau_2)\, d\tau_1 d\tau_2}{\int_{0}^{\xi_{p_1}} D_{\tau_1}(\tau_1)\, d\tau_1} = \frac{p(1-p) + q(1-p_2)}{p + q - pq}. \tag{10}$$

Comparing (10) to (7) we obtain that $P\left(\tau_2 \geq \xi_{p_2} \mid \tau_1 \leq \xi_{p1}\right)$ is greater than $P\left(\tau_2 \geq \xi_{p_1} \mid \tau_1 \leq \xi_{p1}\right)$, which implies that testing the second hypothesis in the sequence using the same critical value as for testing the first hypothesis, provides a lower bound to the probability of retaining the relevant variable using the correct critical values.

The following derivations assume constant critical values to illustrate that the results in Section 3.1 hold when the test-statistics have otherwise different distributions. We consider models with three and four regressors, two of which are relevant.

3.2.2 k regressors: two relevant variables

We wish to examine first whether the probability of retaining relevant variables is larger in more general models. To do so notice that the joint density of the ordered t^2-statistics in a model with k regressors when only two of those are relevant is:

$$D_{\tau_1,\ldots,\tau_k}(\tau_1,\ldots,\tau_k) = \sum_{\mathcal{I}} \prod_{i} f(\tau_{j_i})\, g_1(\tau_{j_m})\, g_2(\tau_{j_n})$$

for $0 \leq \tau_1 \leq \cdots \leq \tau_k < \infty$, and zero otherwise. $g_1(\cdot)$ and $g_2(\cdot)$ are densities of non-central distributions, $\prod_i$ is the product of $k-2$ central densities $f(\cdot)$, and $\sum_{\mathcal{I}}$ denotes the summation over all terms in which $j_1, \ldots, j_{k-2}, j_m, j_n$ are the k elements of the $k!$ permutations of $1, \ldots, k$. So, denoting $p = F\left(\xi_p\right)$, $q_1 = G_1\left(\xi_p\right)$ and $q_2 = G_2\left(\xi_p\right)$:

$$P\left(\tau_k > \xi_p \mid \tau_{k-1} \leq \xi_p\right) = 1 - \frac{p q_1 q_2}{p\delta + (k-2)\, q_1 q_2 \theta}. \tag{11}$$

and:

$$\begin{aligned} & \mathsf{P}\left[\tau_{k-1} > \xi_p \mid \tau_{k-2} \leq \xi_p\right] \\ = \; & 1 - \frac{p^2\delta + (k-2)\, pq_1q_2\theta}{p^2 + (k-2)\, p\delta\theta + \frac{1}{2}(k-2)(k-3)\, q_1q_2\theta^2} \end{aligned} \tag{12}$$

which converge to unity as k increases. In addition, (11) is larger than (7) if non-central distributions imply $q_2 < p$, as it seems to be the case for the χ^2 and $F-$distribution. Hence, for fixed k, more relevant variables do not lead to lower probabilities of correctly including the most significant variable. Finally, (11) is larger than (12) which indicates that sequential testing does not reduce the probability of finding the DGP.

3.2.3 Compare to Hoover and Perez (1999)

To illustrate we next compare the computed probabilities from (11) and (12) with the frequencies of retaining relevant variables in Hoover and Perez (1999) DGPs 6, 6A and 6B. In all those three DGPs variable x_{11} has $\psi = 187$ and $\mathsf{E}[t_0^2] = 47.3$. So, columns 2-3 in our Table 5 report that information for variable x_3 only. However, for comparison, columns 4-5 provide the computed probabilities, and the frequencies of retaining both variables in HP's Table 5 for DGPs 6A and 6B, and their Table 7 for DGP 6.

Testing is carried out at the 1% significance level for DGP 6, and at 5% for DGPs 6A and 6B. The approximate critical value is close to 8 and $\mathsf{E}[t_0^2] > 40$ for variable $x11$ in all DGPs, and for variable $x3$ in DGP 6B. So, we would expect a large probability of keeping those relevant variables. We have found that the probabilities of retaining x_{11} and x_3 (see column 4) are unity, in agreement with HP's frequencies reported in our column 5. For variable $x3$ in DGPs 6 and 6A $\mathsf{E}[t_0^2]$ is 0.87 and 12, so we expect the probability of retaining $x3$ to be very low in DGP 6 and much larger in DGP 6A. Table 5 indicates that this variable is to be kept only 5 out of 100 times for DGP 6, and almost always (99.9 times out of 100) for DGP 6A. The corresponding HP frequencies of retention can be read out of Column 5 and are only 0.8% and as large as 86.4%, respectively.

$\mathcal{DGP}$	ψ	$\mathsf{E}[\mathsf{t}_0^2]$	Prob.		$\mathcal{HP}$	
	$x3$	$x3$	$x3$	$x11$	$x3$	$x11$
$y6 = 0.67x11 - 0.023x3 + \varepsilon$	0.24	0.87	0.054	1.000	0.008	0.998
$y6A = 0.67x11 - 0.32x3 + \varepsilon$	46	12.0	0.999	1.000	0.864	0.999
$y6B = 0.67x11 - 0.65x3 + \varepsilon$	190	47.9	1.000	1.000	0.996	0.997

Table 5: HP's DGPs, non-centrality parameters, mean values of t-squared, computed probabilities of retention, and HP's frequencies of retention.

4 Conclusion

Repeated testing and generality of the initial model do not seem to have the commonly believed pernicious effect on the probabilities of correct inclusion, when the initial model contains relevant variables and the selection procedure incorporates ordering of variables according to their squared $t-$values. The following results are found above: (i) the probability of choosing the relevant variables increases with the total number of variables in the initial model; (ii) the chance of including least relevant variables is not larger than that of keeping the most relevant variables; (iii) the probability of correct inclusion increases with the proportion of relevant variables in the initial model; and (iv) incorrect retention is less likely than correct inclusion but it is not smaller when the relevant variables are more important. Those results are derived assuming that t^2-statistics are mutually independent, and that those associated to the relevant variables have the same non-central distribution. However, relaxing the assumption of identical distribution does not seem to affect those results.

The story when all variables in the initial model are irrelevant is as follows: (i) the probability of correct exclusion is not reduced by sequential testing, it is larger when testing at lower significance levels, and it is smaller for more general initial models; (ii) the probability of incorrect inclusion is enhanced by sequential testing, it is sensibly reduced when testing at smaller significance levels, and increases with model generality; and (iii) the benefit from sequential testing more than offsets the evil of incorrect inclusion since the probabilities of correct exclusion are larger than the probabilities of incorrect inclusion.

References

Campos, J., D.F. Hendry, and H.M. Krolzig (2003). Consistent model selection by an automatic Gets approach. mimeo.

David, H.A. (1981). Order Statistics. New york: J. Wiley and Sons, Inc.

Hansen, B. (1999). Discussion of data mining reconsidered. Econometrics Journal 2, 192–201.

Hendry, D.F. and H.M. Krolzig (1999). Improving on 'data mining reconsidered' by K.D. Hoover and S.J. Perez. Econometrics Journal 2, 202–219.

Hendry, D.F. and H.M. Krolzig (2001). Automatic Econometric Model Selection. London: Timberlake Consultant Press.

Hogg, R.V. and A.T. Craig (1970). Introduction to Mathematical Statistics. London: The MacMillan Company.

Hoover, K.D. and S.J. Perez (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. Econometrics Journal 2, 167–191.

Lovell, M.C. (1983). Data mining. The Review of Economics and Statistics 65, 1–12.