Specification Search and Stability Analysis

J. del Hoyo

J. Guillermo Llorente [1]

Universidad Autónoma de Madrid

This version: May 10, 1999

[1]We are grateful for helpful comments to Richard Watt, and seminar participants at Dept. de Economía Cuantitativa, Universidad Autónoma de Madrid. We thank Ana Rubio for excellent computer assistantship. Partial financial support from the DGICYT under grant PB94-018 is acknowledged. Correspondence: Dept. de Economía Cuantitativa, Universidad Autónoma de Madrid, 28049 Madrid, SPAIN. Phone: +34-91-397-4812 or +34-91-397-5033. Fax +34-91-397-4091. E-mail: juan.hoyo@uam.es or guiller@uam.es.

Abstract

Most specification search processes select models based on some goodness of fit statistic (i.e. R^2 or related F). The effects of the sequential search on the statistical tests should be taken into account when looking for the maximum goodness of fit. To avoid misspecified models it is useful to study the selected models based on the full sample, and along the sample. This paper presents a conditional sequential procedure to be used in the specification search process of linear regression models, as a way to minimize data snooping or data mining. It is a combined test, first it considers the search for the "best" set of regressors, and conditional on this set, it studies their significance along the sample. The characteristics of the conditional test are presented. Its usefulness is considered with one application.

Key Words: Specification Search, R^2, $R^2_{\max}$, Rolling tests, Statistical significance and Stability Analysis, Brownian Motion, Monte Carlo Simulations.

1 Introduction

Specification analysis is an important step in selecting a model for either structural analysis or forecasting. To explain a given variable, one must choose the optimal subset of k predictors among a set of m indicated variables proposed by the relevant theory or previous studies. The number k is fixed a priori. This search is usually achieved by maximizing the goodness of fit or R^2 (or its equivalent F). Conducting inference without properly considering such a search or selection process one has data mining that can be extremely misleading.

Foster *et al.* (1997) study model selection using maximum R^2 to get the best regression model having fixed the regresand and k of m regressors. They argue that determining the proper cut-off points of the R^2 distribution requires the researcher to consider the selection procedure, and hence the distribution function of the *maximal* R^2 (hereafter R^2_{max}) has to be used. This function has a difficult form that must either be simulated with Monte Carlo or approximated as in Foster *et al.* (1997) with Bonferroni or Rencher and Pun bounds. White (1997) proposes model selection using a "Reality Check" comparing the forecasting performance of the candidate against a benchmark model. Out-of-sample prediction is a good performance test, but the choice of the benchmark model could be difficult.

Surprisingly the information embodied in the sample is not usually exploited when testing for data mining. Here we argue that if the selected model is tested with both full sample and rolling estimation along the sample, data mining problems should be reduced. Specifically, before accepting a model with significant global R^2_{max} or R^2, it is of value to test whether hypotheses of significant and constant coefficients are true along the full sample. A sound theoretical model should remain valid if estimated and tested along the sample. Foster *et al.* (1997) test for data mining using R^2_{max} estimated with the full sample. It is, however, possible to have models that comply with R^2_{max} statistics while being spurious (nonconstant and/or non significant coefficients). In this paper we propose to consider the information from the rolling estimations to detect this situation.

This paper adds to process of model selection and data mining the idea that model parameters may not be significant or change through time, which can bias the choice of the benchmark model or the specification search among the m variables. The presence of non

significant and/or time-varying parameters (TVP) when they are assumed constant is a sign of misspecification error, possibly contaminating subsequent analyses. Models selected only on the basis of the full sample could be misleading and can increase the potential data mining problems. It is in this context that del Hoyo and Llorente (1998) study the improvement in forecasting considering non constant parameters. Here we consider a sequential procedure for both means (discrimination and stability) to decrease biases in choosing a model. The first stage uses the R^2 or R^2_{max} as a discrimination procedure to select the optimal number of regressors or explanatory variables. The second stage tests significance of this relationship by means of the rolling statistics[1]. The conditional distributions of the rolling statistics are tabulated, conditional on the discrimination stage. The innovation here is the sequential consideration of both procedures.

The tests are studied for two sample sizes, 500 and 1000 observations, considered to be representative of empirical work in economics. The maximum value for m is 10. The rolling statistics are studied for three windows determined as a fix proportion of the sample ($\lambda_0 = 1/4,\ 1/2,$ and $3/4$). We find that the degree of dependence is important. It increases with the rejection level of both tests. Thus the tabulation of the conditional distributions should be considered. We conclude about the less likelihood of accepting spurios models when using the combined proposed procedure.

The paper proceeds as follows. Section 2 presents and tabulates the distributions of the relevant statistics. Section 3 introduces the sequential procedure described above. The conditional distributions are studied. Section 4 gives an illustration with a model proposed by Campbell, Grossman and Wang (1993). Section 5 concludes.

2 Discrimination and Rolling Tests

This section reviews the discrimination tests (variable selection statistics), in particular the R^2_{max}, and the rolling tests.

Assume the model under consideration is an asset pricing model, where y is a $T \times 1$ vector

[1]Hoyo and Llorente (1999) studies the same problem but using recursive statistics.

of security returns. The potential predictor variables of these returns are m variables, and the researcher chooses k out of m ($k \leq m$). The number of potential regressors (m) does not have any limit a priori. The number of regressors (k) depends on the problem at hand (it can even include lagged returns), but it is often less than ten and is fixed a priori.

To emphasize the predictive nature of the considered model, and for easy exposition motives, the observations on y_t are assumed to be generated by a model of the following type[2]

$$y_t = \boldsymbol{X_{t-1}}\boldsymbol{\beta_t} + \epsilon_t \,, \qquad t = 2, 3, \ldots, T. \tag{1}$$

Under the null hypothesis $\boldsymbol{\beta_t} = \boldsymbol{\beta}$ is a $(k+1)\times 1$ vector of k constant slope parameters plus the intercept, the errors (ϵ_t) are assumed to be a martingale difference sequence with respect to the σ-fields generated by $\{\epsilon_{t-1}, \boldsymbol{X_{t-1}}, \epsilon_{t-2}, \boldsymbol{X_{t-2}}, \ldots\}$, and $\boldsymbol{X_t}$ is a $1 \times (k+1)$ vector of regressors. The regressors are assumed to be constant and/or $I(0)$ with $E(\boldsymbol{X_t'X_t}) = \boldsymbol{\Sigma_X}$. Denote $\lambda = \frac{t}{T}$, $\Rightarrow$ weak convergence on $D[0,1]$, and $[\bullet]$ the integer part of the value inside brackets. Also assume that $T^{-1}\sum_{i=1}^{[T\lambda]} \boldsymbol{X_i'X_i} \xrightarrow{P} \lambda\boldsymbol{\Sigma_X}$ uniformly in λ for $\lambda \in [0,1]$; $E(\epsilon_t^2) = \sigma^2 \;\forall\; t$, and $T^{-1}\sum_{i=2}^{[T\lambda]} \boldsymbol{X_i'}\epsilon_i \Rightarrow \sigma\boldsymbol{\Sigma^{1/2}}\boldsymbol{W_k}(\lambda)$, with $\boldsymbol{W_k}(\lambda)$ a k-dimensional vector of independent Wiener or Brownian motion processes[3]. $\boldsymbol{X_{t-1}}$ can include lags of the dependent variable as long as they are $I(0)$ under the null (see Stock 1994). Given the m candidates explanatory variables, there are $N = \begin{pmatrix} m \\ k \end{pmatrix}$ possible model regression specifications with k explanatory variables.

2.1 Distribution of the Maximal R^2

Under the classical assumptions, the R^2 of the regression, representing the proportion of variation in the dependent variable explained by the regressors, is distributed as Beta $\left(\frac{k}{2}, \frac{T-(k+1)}{2}\right)$, under the null hypothesis that $\boldsymbol{\beta}_{k\times 1} = \mathbf{0}$ (all the slope coefficients of the linear regression are equal to zero, against the alternative that at least one of the coefficients is different from zero)[4] [5]. This distribution does not take into consideration the selection process among

[2] Boldface is used for vectors and matrices.

[3] ϵ_t can be conditionally (on lagged ϵ_t and $\boldsymbol{X_t}$) heteroskedastic and the results do not change.

[4] The intercept is excluded from the hypothesis.

[5] The significance of the regression can also be tested using the $F-$statistic.

the potential explanatory variables. Thus, the cutoff values of the distribution for the R^2 statistics need to be adjusted for this process[6], assuming the "best" k regressors have been chosen by maximizing the R^2.

The distribution function for the R^2_{max} may be derived applying the standard order statistic argument. Nevertheless, for the considered model, where y does not change among regressions, the selection process induces overlapping elements in the matrices, and there may exist correlation among regressors, this is a difficult task. This induces the use of some approximation[7] for the joint distribution function of the R^2_{max}. In particular, the Bonferroni bound is $U_{R^2}(r) \geq 1 - \{[1 - \text{Beta}(r)]\, N\}$, and the Rencher and Pun (1980) approximation to the cutoff levels is given by $R^2_\gamma \approx F^{-1}\left[1 + (\ln(\gamma)/\ln(N)^{1.8N^{0.04}})\right]$, where γ is the percent cutoff level, and F^{-1} is the inverse of the beta cumulative distribution function.

The distribution for the R^2_{max} can also be computed numerically using Monte Carlo simulations. In this paper this distribution as well as the distributions related to other statistics will be numerically computed. Table 1 illustrates the differences between the cutoff levels computed with the approximated bounds (values taken from Foster *et al.* 1997), and those calculated numerically by simulations. Comparing these values it can be seen that the Bonferroni bound is very conservative, and that the Rencher and Pun approximation is close to the numerical results.

2.2 Rolling Tests

This section presents the asymptotic distributions for the rolling statistics that consider the possibility of at least one change in the parameters along the sample with unknown a priori break date. Rolling tests can be considered as a tool to detect misspecifications, in particular, spurious models. It is assumed that the optimal set of regressors (k) is already chosen. Rolling statistics are computed using subsamples of constant size $s = [\lambda_0 T]$ along the sample. From $t_* = [T(\lambda - \lambda_0)] + 1, \ldots, [T\lambda]$, which are equivalent to $t = s, \ldots, T$.

[6] The cutoff values for the $F-$statistic should also be adjusted.
[7] See Foster *et al.*, 1997.

Table 1: Maximal R^2 cutoff levels: comparisons between methods

Number of Potential Regressors (m)	Number of Regressors Selected (k)				
	1	2	3	4	5
Panel A: Bonferroni Bound					
10	0.036	0.055	0.071	0.084	0.094
25	0.040	0.068	0.094	0.116	0.136
50	0.044	0.079	0.110	0.138	0.164
100	0.048	0.089	0.126	0.159	0.189
Panel B: Rencher/Pun Rule-of-Thumb					
10	0.027	0.046	0.060	0.071	0.079
25	0.032	0.054	0.072	0.088	0.103
50	0.035	0.060	0.081	0.100	0.119
100	0.038	0.066	0.090	0.113	0.135
Panel C: Numerical Calculations					
10	0.031	0.045	0.056	0.062	0.066
25	0.037	0.058	0.073	0.085	0.095
50	0.042	0.065	0.084	0.100	0.114
100	0.043	0.067	0.088	0.106	0.121

NOTE: Entries are the 95 percent cutoff values for the "best" k-variable regressions R^2 given different number of potential regressors (m) and a fixed sample size of 250 observations. The table reports the 95 percent confidence limits for R^2 for the null hypothesis that all of the slope coefficientes of and OLS regression are equal to zero, where (a) only k of m potential regressors are used, (b) all possible regression combinations are tried, and (c) only the regression with the highest R^2 is reported. The alternative hypothesis is that at least one of the OLS slope coefficients is not equal to zero. The Bonferroni inequality is a bound and therefore represents a conservative test. The Rencher/Pun rule-of-thumb is an approximation of the exact distribution. The Numerical Calculations are computed by Monte Carlo simulations based on 5000 replications. The values in Panel A and B are from Foster *et al.* (1997) Table II, pp. 599.

The rolling OLS coefficients can be written as random elements on $D[0,1]$ as

$$\hat{\boldsymbol{\beta}}(\lambda;\lambda_0) = \left(\sum_{t_*}^{[T\lambda]} \boldsymbol{X}'_{\boldsymbol{t-1}}\boldsymbol{X}_{\boldsymbol{t-1}}\right)^{-1}\left(\sum_{t_*}^{[T\lambda]} \boldsymbol{X}'_{\boldsymbol{t-1}}y_t\right) \qquad 0 \le \lambda_0 \le \lambda \le 1. \tag{2}$$

The Wald type statistics used to test $q \le (k+1)$ linear independent restrictions on $\boldsymbol{\beta}$ ($H_0 : \boldsymbol{R}\hat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{r}$, where $\boldsymbol{R}$ is a $q \times (k+1)$ nonstochastic matrix, and $\boldsymbol{r}$ is a $q \times 1$ vector) have as general form

$$\hat{F}_T(\lambda;\lambda_0) = \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}(\lambda;\lambda_0) - \boldsymbol{r}\right)'\left[\boldsymbol{R}\left(\sum_{t_*}^{[T\lambda]} \boldsymbol{X}'_{\boldsymbol{t-1}}\boldsymbol{X}_{\boldsymbol{t-1}}\right)^{-1}\boldsymbol{R}'\right]^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}}(\lambda;\lambda_0) - \boldsymbol{r}\right)}{q\hat{\sigma}^2(\lambda;\lambda_0)}, \tag{3}$$

where $\hat{\sigma}^2(\lambda;\lambda_0)$ is the rolling estimation of the variance, and t_* the corresponding beginning of the rolling computations.

The asymptotic behavior of this statistic is derived by applying the Functional Central Limit Theorem (FCLT) and the Continuous Mapping Theorem (CMT). Therefore, $\hat{F}_T(\lambda;\lambda_0) \Rightarrow \hat{F}(\lambda;\lambda_0)$. The form of the final distributions depends on $\boldsymbol{R}$ and $\boldsymbol{r}$ as determined by the null hypothesis.

In what follows, the subscripts refer to the null hypothesis to be tested, and the superscripts to the number of parameters involved. For example, $\hat{F}(\lambda;\lambda_0)_0^k$ represents the particularization of the statistic to test the null hypothesis that all the slope parameters of the model rolling estimated are equal to zero.

Following Stock (1994), the Wald type statistic to test for the significance of q coefficients along the sample ($H_0 : \boldsymbol{R}\hat{\boldsymbol{\beta}}(\lambda;\lambda_0) = \boldsymbol{0}$), has the following asymptotic distribution

$$\hat{F}(\lambda;\lambda_0)_0^q \equiv \max_{\lambda_0 \leq \lambda \leq 1} \left[\lambda_0 q * \hat{F}(\lambda;\lambda_0)\right] \Rightarrow \sup_{\lambda_0 \leq \lambda \leq 1} \boldsymbol{H_q}(\lambda;\lambda_0)'\boldsymbol{H_q}(\lambda;\lambda_0), \tag{4}$$

where $\boldsymbol{H_q}(\lambda;\lambda_0) = \boldsymbol{W_q}(\lambda) - \boldsymbol{W_q}(\lambda - \lambda_0)$.

The statistic to test for the stability of the model along the sample, comparing the recursive estimations with those from the full sample, has the following expression (H_0 : $\boldsymbol{R}\{\hat{\boldsymbol{\beta}}(\lambda;\lambda_0) - \hat{\boldsymbol{\beta}}(1)\} = \boldsymbol{0},\ \lambda \in [\lambda_{min}, 1]$)[8]

$$\hat{F}(\lambda;\lambda_0)^q_{\hat{\boldsymbol{\beta}}(1)} \equiv \max_{\lambda_0 \leq \lambda \leq 1} \left[\lambda_0 q * \hat{F}(\lambda;\lambda_0)\right] \Rightarrow \sup_{\lambda_0 \leq \lambda \leq 1} \boldsymbol{G_q}(\lambda;\lambda_0)'\boldsymbol{G_q}(\lambda;\lambda_0), \tag{5}$$

where $\boldsymbol{G_q}(\lambda;\lambda_0) = \boldsymbol{W_q}(\lambda) - \boldsymbol{W_q}(\lambda - \lambda_0) - \lambda_0 \boldsymbol{W_q}(1)$.

2.3 Monte Carlo Results

Critical values for the R^2_{max} and the rolling statistic is reported in this section. The basic model is of the type presented in Equation (1) corresponding to several possible choices of k.

Following the convention established before, the rolling statistic to be studied in this paper is denoted by $\hat{F}(\lambda;\lambda_0)_0^k$, to test the statistical significance of all parameters except

[8] Notice that $\hat{\boldsymbol{\beta}}(1;1)) = \hat{\boldsymbol{\beta}}(1))$ is the estimator with the full sample.

Table 2: R^2_{max} cutoff levels

Percentile	Number of Regressors Selected (k)									
	1	2	3	4	5	6	7	8	9	10
Panel A: 500 observations										
0.100	0.013	0.019	0.024	0.027	0.029	0.030	0.031	0.031	0.032	0.032
0.050	0.015	0.023	0.027	0.031	0.033	0.034	0.035	0.036	0.036	0.037
0.001	0.022	0.030	0.035	0.039	0.041	0.043	0.044	0.045	0.045	0.046
Panel B: 1000 observations										
0.100	0.006	0.010	0.012	0.014	0.015	0.015	0.016	0.016	0.016	0.016
0.050	0.008	0.012	0.014	0.016	0.017	0.017	0.017	0.018	0.018	0.018
0.001	0.011	0.016	0.018	0.019	0.021	0.022	0.022	0.023	0.023	0.023

NOTE: Entries are the 90, 95 and 99 percent cutoff values for the "best" k-variable regressions R^2 given $m = 10$ potential regressors (m) and two sample sizes 500 and 1000 observations. The table reports the confidence limits for R^2 for the null hypothesis that all of the slope coefficients of and OLS regression are equal to zero where (a) only k of $m = 10$ potential regressors are used, (b) all possible regression combinations are tried, and (c) only the regression with the highest R^2 is reported. The alternative hypothesis is that at least one of the OLS slope coefficients is not equal to zero. The numerical calculations are computed by Monte Carlo simulations based on 10000 replications.

the intercept. The intercept is not considered in the statistics to keep in line with the usual tests, and with the R^2_{max}.

Table 2 presents the tabulation for the R^2_{max}. The cutoff levels are 90, 95 and 99 percent and $T = 500$ and 1000 observations. The regressions depend on the number of regressors chosen (k) between the potential independent variables ($m = 10$). The numerical distributions where calculated using 10000 replications, and the m regressors where simulated as independent $N(0, 1)$ variables.

Approximate critical values for the rolling statistics are reported in Table 3. Entries are the *sup* values of the functionals of Brownian motions. The critical values were computed performing Monte Carlo simulations of the limiting functionals of Brownian Motion processes involved in the statistics. All critical values were computed using 10000 replications and $T = 3600$ in each replication.

Table 3: Critical Values Rolling Tests

Percentile	Number of Regressors k									
	1	2	3	4	5	6	7	8	9	10
$F(\lambda;\lambda_0)_0^k$, $\lambda_0 = 1/4$										
0.100	8.383	5.696	4.636	4.041	3.626	3.315	3.102	2.929	2.794	2.690
0.050	9.928	6.562	5.224	4.501	3.999	3.664	3.402	3.202	3.040	2.928
0.001	13.406	8.534	6.526	5.407	4.840	4.427	4.063	3.775	3.600	3.388
$F(\lambda;\lambda_0)_0^k$, $\lambda_0 = 1/2$										
0.100	6.402	4.508	3.702	3.308	3.027	2.815	2.665	2.547	2.423	2.362
0.050	7.833	5.310	4.344	3.835	3.448	3.176	2.983	2.820	2.708	2.587
0.001	11.127	7.218	5.844	4.912	4.401	3.963	3.671	3.427	3.259	3.054
$F(\lambda;\lambda_0)_0^k$, $\lambda_0 = 3/4$										
0.100	4.846	3.636	3.065	2.765	2.555	2.400	2.308	2.209	2.134	2.069
0.050	6.325	4.396	3.591	3.238	2.969	2.803	2.652	2.491	2.394	2.309
0.001	9.542	6.241	4.863	4.188	3.783	3.554	3.323	3.137	2.942	2.797

NOTE: Entries are the *sup* values of the functionals of Brownian motion. All critical values were computed by Monte Carlo simulation of the limiting functionals of Brownian motion as described in the main text. They are based on 10000 Monte Carlo replications and T=3600 observations. k is the number of parameters in the regression excluding the intercept, λ_0 the size of the rolling window as a proportion of the full sample, and $\hat{F}(\lambda;\lambda_0)_0^k$ the statistic tests for the null hypothesis that all coefficients in the regression but the intercept are equal to zero.

3 Sequential Specification Procedure

The methodologies presented in the last section have been used in previous papers as useful tools in specification search processes. Nevertheless, they have not been studied when applied sequentially. It seems natural to consider both tests to reduce the likelihood of data snooping or data mining. We could find models that satisfy the R^2_{max} criterion but their coefficients are not significant along the sample. This could be due to the fact that the R^2 concentrates on the residual sum of squares calculated from the full sample, but it does not take into consideration possible misspecifications. If the misspecification shows up as non significant or time varying parameters, they can be detected with tests based on rolling estimators.

This paper proposes a sequential specification procedure in two steps. First, the procedure selects one model based on either the R^2 or the R^2_{max} (if there has been selection process among regressors). Once the model is accepted, the second step is to apply the rolling statistics. Rolling statistics can test the significativeness of the coefficients along the sample, and/or test the statistical discrepancy between the rolling and the full sample estimates. The nature of the sequential specification procedure makes it necessary to calculate

the conditional distributions of the recursive statistics conditional to the first step.

Before studying the conditional distributions we would like to point that the R^2 has to be used in those ideal situations where the researcher has not done sequential selection among variables. We will concentrate mainly on the conditional distribution to test for rolling statistical significance. The distribution that conditions will be either the R^2 or the R^2_{max} ($F(\lambda;\lambda_0)^k_0|R^2$, or $F(\lambda;\lambda_0)^k_0|R^2_{max}$)[9]. Given the analytical difficulties, the recursive conditional distributions presented in this paper were simulated by Monte Carlo methods. The simulation experiment is described below.

The objective is to derive the conditional probability function for the rolling statistics depending on the $(1-\alpha_1)$ probability chosen for the R^2 or R^2_{max}. To achieve this objective we simulated 10000 models under the null hypothesis for the R^2 or R^2_{max} (the null is that all slope parameters are zero), the corresponding R^2 or R^2_{max} were calculated to obtain its distribution later. For each of these models the $F(\lambda;\lambda_0)^k_0$ statistic was also calculated. The next step was to select only the $F(\lambda;\lambda_0)^k_0$ values corresponding to those models that comply with the chosen $(1-\alpha_1)$ probability for the R^2 or R^2_{max}. These $F(\lambda;\lambda_0)^k_0$ values form the conditional distribution. The result can be presented in a three dimensional graph. It is a probability surface. The shape of the surface gives an idea about the characteristics of the conditional distributions. The experiment was done for $m = 10$, $k = 1, 3, 5$, and 7, sample sizes $T = 500$ and 1000 observations respectively, and for three window sizes $\lambda_0 = 1/4,\ 1/2,$ and $3/4$. We think these situations are representative enough to ascertain the characteristics of the conditional distributions.

Figure 1 presents the probability surface for the combined test $F(\lambda;\lambda_0)^k_0|R^2_{max}$ for $k = 3$, $T = 1000$ observations, and $\lambda_0 = 1/2$ (500 observations rolling window). The x-axis presents the $(1-\alpha_2)$ probability for the F statistic that will be used for the conditional distribution, the y-axis shows the $(1-\alpha_1)$ probability for the R^2_{max} statistics, and the z-axis gives the values for $F(\lambda;\lambda_0)^k_0|R^2_{max}$. The way to interpret the graph is the following. Choose a $(1-\alpha_1)$ probability for the R^2_{max}, the corresponding F-curve for the conditional distribution

[9] Following a similar procedure it is also possible to tabulate the conditional distributions for $F(\lambda;\lambda_0)^k_{\hat{\boldsymbol{\beta}}_{(1)}}|R^2$, or $F(\lambda;\lambda_0)^k_{\hat{\boldsymbol{\beta}}_{(1)}}|R^2_{max}$.

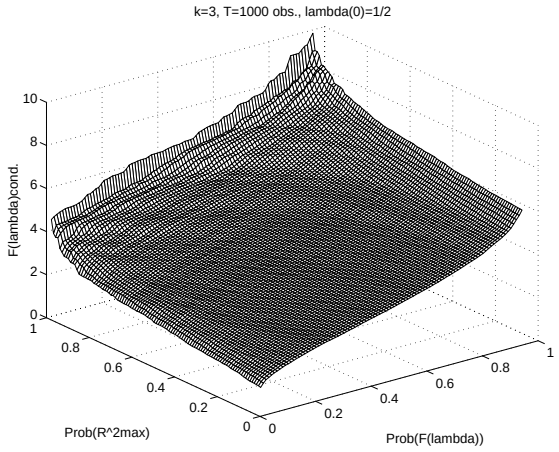

Figure 1: Conditional distribution $F(\lambda;\lambda_0)_0^k|R^2_{max}$

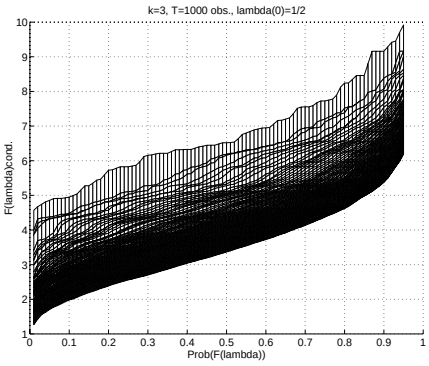

Figure 2: Conditional distribution $F(\lambda;\lambda_0)_0^k|R^2_{max}$

is determined by the intersection between the probability surface and the $y = (1-\alpha_1)$ plane. The curve for a $Prob(R^2_{max}) = 0$ (i.e. when H_0 is rejected in all the models) corresponds to the unconditional $F(\lambda;\lambda_0)_0^k$.

The probability surface is quite smooth but at the extremes, particularly along the y-direction. Its steepness increases with the probability along both axes. Figure 2 presents the same surface but from a different view point. Now the graph is viewed from the x-axis side (it is represented in two dimensions), the result is an envelop of curves. The higher the dependence between both tests the wider the envelop. The degree of dependence is evident when observing both graphs[10].

[10]The graphs not presented, corresponding to the other studied combinations, have similar characteristics.

The main conclusion from the combined conditional test is that the degree of dependence between both tests increases with the probability of accepting H_0. Thus, there is an important loss of efficiency by using the unconditional distributions instead of the conditional ones. To get an idea of the loss of efficiency Table 4 compares unconditional and conditional values for the rolling statistic, for $T = 1000$ observations, $\lambda_0 = 1/2$, $m = 10$ and $k = 1,\ 3,\ 5,\ 7$. Entries are calculed following the same steps as those explained above for the sequential procedure[11].

Table 4: Distributions R^2_{max}, $F(\lambda;\lambda_0)$, and $F(\lambda;\lambda_0)|R^2_{max}$, for $n = 1000, \lambda_0 = 1/2, m = 10$

Percentile	k	R^2_{max}	$F(\lambda;\lambda_0)$	$F(\lambda;\lambda_0)\|R^2_{max} = 90\%$	$F(\lambda;\lambda_0)\|R^2_{max} = 95\%$	$F(\lambda;\lambda_0)\|R^2_{max} = 99\%$
	1	0.006	10.349	14.746	16.115	18.827
0.10	3	0.012	5.368	7.377	7.969	9.159
	5	0.014	3.902	5.050	5.389	5.793
	7	0.015	3.096	4.028	4.110	4.570
	1	0.008	11.996	16.681	18.753	20.219
0.05	3	0.014	6.191	8.294	8.632	9.919
	5	0.016	4.419	5.462	5.565	6.045
	7	0.018	3.433	4.277	4.556	4.909
	1	0.011	15.895	21.499	21.993	25.351
0.01	3	0.017	7.775	9.919	10.756	10.760
	5	0.021	5.331	6.061	6.699	7.432
	7	0.022	4.105	5.168	5.214	6.622

NOTE: Entries are critical values computed by Monte Carlo simulation as described in the main text, Section 3. They are based on 10000 Monte Carlo replications, $T = 1000$ observations, $\lambda_0 = 1/2$, $m = 10$, and $k = 1,\ 3,5,$ and 7. m is the number of potential regressor, k is the number of parameters (regressors) considered in the regression excluding the intercept, λ_0 the size of the rolling window as a proportion of the full sample, R^2_{max} the maximal goodness of fit statistic, and $\hat{F}(\lambda;\lambda_0)^k_0$ the rolling statistic to test for the null hypothesis that all coefficients in the regression but the intercept are equal to zero.

The third column in Table 4 presents the tabulated values corresponding to the R^2_{max}. The fourth column contains the unconditional tabulated values for the rolling statistic. These values are slightly higher than those in Table 3 because they consider the selection process, though in an unconditional way. The last three columns present the conditional values for the rolling statistic conditonal upon different levels of acceptance for the null hypothesis of the R^2_{max}. Comparing the last four columns in Table 4 we see the increasing values for the rolling

[11]The other tables not presented have similar characteristics.

statistic as the acceptance level of the R^2_{max} increases. Thus, as $F(\lambda;\lambda_0)|R^2_{max} > F(\lambda;\lambda_0)$ the number of models with non significant rolling coefficients increases. This means, less spurios models are going to be accepted. Therefore, the main hypothesis about the decrease in the likelihood of data snooping or data mining is confirmed. As a rule of thumb, we observe in all the experiments, that the values for the unconditonal rolling statistics at the 1% percentile is very close to the conditonal value at the 10% percentile conditonal upon $R^2_{max} = 90\%$. These unconditonal values can be used as bounds for the true ones.

4 Empirical Application

This section presents one example using the previously proposed sequential procedure. The model is postulated in Campbell, Grossman and Wang (1993). The stability of the example is studied in Hoyo and Llorente (1998). The application is done for the USA and Spanish stock markets. The data is composed by daily observations, with sample periods 7/3/62–12/31/93 for USA, and 1/4/92–12/31/95 for Spain. The working expression is the following equation

$$R_{t+1} = \beta_0 + (\sum_{i=1}^{5} \beta_i D_i + \gamma V_t) R_t + \epsilon_t, \tag{6}$$

where R_t is the return of the market on day t, V_t is the volume traded on the market on day t, and D_i is a dummy variable corresponding to the $i-th$ day of the week. The consideration of dummy variables, one for each day of the week, tries to account for the accepted different behavior of the relationship between returns depending on the day of the week. The characteristics of small market, and heavy external influence on the Spanish market, leads us to consider the foreign influence in the form of two additional variables representing the USA stock market behavior. The first one is called *dowov* and represents the overnight return of the Dow Jones Index; the second one called *dowin* reports the intraday return on the Dow Jones Index. Both variables are included in the equation corresponding to the Spanish market.

Table 5: Maximal R^2 cutoff levels

Percentile	$m = 14, k = 7$	$m = 16, k = 9$
0.100	0.0024	0.0219
0.050	0.0027	0.0244
0.001	0.0037	0.0289

Campbell *et al.* (1993) [12] try 14 potential explanatory variables, and conclude that the best regression is the one presented in Equation 6, with 7 predictor variables for the USA data, for the Spanish data two more variables where added as explained above. Thus, the R^2_{max} should be tabulated for $m = 14, k = 7, T = 8000$ for the USA, and for $m = 16, k = 9, T = 1000$ for the Spanish example. The cutoff levels for the R^2_{max} are presented in Table 5. In Hoyo and Llorente (1998) the estimated R^2 were 0.055 and 0.17 for USA and Spain respectively. Therefore, comparing these values with those in Table 5 we reject the null hypothesis in favor of the alternative about the existence of at least one slope parameter different from zero.

Thus, the model is accepted according to the R^2_{max} statistic. The next step is to validate its significativeness using the conditional distributions for the rolling statistics. Table 6 presents the tabulations corresponding to the characteristics of the chosen example. The tabulations for $T = 8000$ (USA example) were done with $T = 1000$ because of computing problems. Therefore, the values in the table should be considered as bounds for the right ones.

Rolling estimations for the considered data were calculated to test the hypothesis of significance. This test studies the validity of the relationship through time (all the parameters but the intercept equal to zero). The results are summarized in Table 7, entries are test statistics. Tests are significant at the *** 1 percent level, using the tabulations from Table 6.

To summarize, the R^2_{max} statistic provides evidence that the relation is significative at the 1% level for both countries. The statistic $\hat{F}(\lambda; \lambda_0)^k_0 | R^2_{max}$ rejects the null hypothesis that

[12] The example differs slightly from Campbell *et al.* (1993). We assume they have fixed a priori k and m.

Table 6: Critical Values for Rolling Tests

	λ_0	$F(\lambda;\lambda_0)\|R^2_{max}=90\%$	$F(\lambda;\lambda_0)\|R^2_{max}=95\%$	$F(\lambda;\lambda_0)\|R^2_{max}=99\%$
	1/4	4.301	4.373	4.794
$k=7$	1/2	4.405	4.660	5.242
90%	3/4	4.270	4.573	5.304
	1/4	4.686	4.791	5.540
$k=7$	1/2	4.754	5.205	5.558
95%	3/4	4.628	4.846	5.689
	1/4	5.649	5.855	6.225
$k=7$	1/2	5.518	5.709	6.187
99%	3/4	5.377	5.689	5.938
	1/4	3.869	4.042	4.822
$k=9$	1/2	3.943	4.095	4.579
90%	3/4	3.822	4.072	4.586
	1/4	4.158	4.255	4.904
$k=9$	1/2	4.209	4.434	5.105
95%	3/4	4.151	4.377	4.622
	1/4	4.973	4.999	n.a.
$k=9$	1/2	4.769	5.105	n.a.
99%	3/4	4.663	4.911	4.762

NOTE: Entries are calculated following the same steps as in Table 4.

Table 7: Empirical Results: Evidence on Rolling Tests

	$\lambda_0=1/4$	$\lambda_0=1/2$	$\lambda_0=3/4$
$k=7$	51.716***	74.412***	85.091***
$k=9$	15.486***	20.561***	23.279***

the slope coefficients are equal to zero along the full sample at the 1% level for USA ($k=7$) and for Spain ($k=9$). Thus, we conclude accepting the model with the number of included variables among the potential ones. The next step should be to study the stability of the relationship along the sample.

5 Summary

In any applied work, closely related to the phenomenon of data mining or data snooping is the specification search process. To reduce the likelihood of accepting spurios models

this paper proposes a combined test. The combined test consists of two statistics. First, a goodness of fit measure (R^2 or R^2_{max}) that considers the sequential search for the best set of explanatory variables. The second is a rolling statistic to test for the significativeness of the considered model. The unconditional distributions of both tests are studied, as well as the conditional distribution for the combined test. The main result of the paper shows the degree of dependence in the steps of the conditional procedure. There is some loss of efficiency in using the unconditional instead of the conditional distributions for the rolling statistics. Using the conditional distributions we increase the number of spurios models rejected, the conditonal test is more "severe" (has higher cutoff values). An example illustrates the applicability of the proposed procedure.

The directions for further research go in two ways. The first one is related to the statistical properties of the rolling tests. The initial results look promising. The second one, includes as an additional requirement for any model to be stable along the sample. Thus, the rolling statistic to test for the stability of the relationship should be considered.

References

[1] Andrews, D.W.K. "Tests for Parameter Instability and Structural Change with Unknown Change Point". *Econometrica*, 61:821–856, 1993.

[2] Campbell, J., S. Grossman, and J. Wang. "Trading Volume and Serial Correlation in Stock Returns". *Quarterly Journal of Economics*, pages 905–940, November 1993.

[3] Foster, F.D., T. Smith, and R.E. Whaley. "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2". *Journal of Finance*, 52, 2:591–607, June 1997.

[4] Hoyo, J. del, and J.G. Llorente. "Stability Analysis and Forecasting Implications". In A.-P. N. Refenes, A.N. Burgess, and J.E. Moody, editor, *Decision Technologies for Computational Management Science*. Kluwer Academic, London, 1998.

[5] Hoyo, J. del, and J.G. Llorente. "Goodness of Fit, Stability and Data Mining". In Y. Abu-Mostafa, B. LeBaron, A. Lo, and A. Weigend, editor, *Proceedings of Computational Finance 1999, CF99*. MIT Press, Boston, 1999. forthcoming.

[6] Hoyo, J. del, and J.G. Llorente. "Recursive Estimation and Testing of Dynamic Models". *Computational Economics*, 1999.

[7] Stock J.H. "Unit Roots Structural Breaks and Trends". In R.F. Engle, and D.L. MacFadden, editor, *Handbook of Econometrics*, chapter 46, pages 2739–2839. Elsevier Science, 1994.

[8] White, H. "A Reality Check for Data Snooping". *Working paper, UCSD, Dept of Economics*, 1997.