Herwartz, Helmut

Working Paper

A note on model selection in (time series) regression models - General-to-specific or specific-to-general?

Economics working paper / Christian-Albrechts-Universität Kiel, Department of Economics, No. 2007,09

Provided in cooperation with:
Christian-Albrechts-Universität Kiel (CAU)

Suggested citation: Herwartz, Helmut (2007) : A note on model selection in (time series) regression models - General-to-specific or specific-to-general?, Economics working paper / Christian-Albrechts-Universität Kiel, Department of Economics, No. 2007,09, http://hdl.handle.net/10419/22025

A note on model selection in (time series) regression models - General-to-specific or specific-to-general ?

by Helmut Herwartz

Christian-Albrechts-Universität Kiel

Department of Economics

Economics Working Paper

No 2007-09

A note on model selection in (time series) regression models - General-to-specific or specific-to-general ?

Helmut Herwartz*

March 12, 2007

Abstract

The paper provides Monte Carlo evidence on the performance of general-to-specific and specific-to-general selection of explanatory variables in linear (auto)regressions. In small samples the former is markedly inefficient in terms of ex-ante forecasting performance.

Keywords: Model selection, specification testing, Lagrange multiplier tests.
JEL Classification: C22; C51

1 Introduction

Model selection, the extraction of relevant covariates or lags in explaining time series observations, is an important issue in applied econometrics. A widely followed strategy is to start with some general, unrestricted model which is subjected to subsequent reduction. In its current state the LSE approach (Hendry 1993, Hendry and Krolzig 2001) comprises a computer automated procedure (PcGets) of multi step diagnostic testing, subsample modeling, encompassing tests etc, that has given rise to the notion of *testimation*. Putting emphasis on search costs as, for instance, software needs or expertise, or the governance of path dependent model diagnostics, it appears tempting to evaluate the scope of simple specification tests in assisting an empirical analyst. It is the purpose of this note to shed light on the potential of the likely simplest specification strategy, the expansion of a (wrong) small model by means of Lagrange multiplier (LM) specification tests (Godfrey 1988). The analysis is focussed on Monte Carlo exercises for both a dynamic and a static data generating process (DGP). Relative merits of alternative modeling strategies are evaluated in terms of ex-ante forecasting efficiency. The next section sketches the alternative avenues of model selection. The Monte Carlo design and results are provided in Section 3. Section 4 concludes. If not stated otherwise, particular concepts of univariate time series modeling are taken from Lütkepohl (2004).

*Institut für Statistik und Ökonometrie, Christian-Albrechts-Universität zu Kiel, Germany, Herwartz@stat-econ.uni-kiel.de

2 Modeling strategies

The issue considered is to choose the 'most informative' variables to explain observations y_t, $i = 1, \ldots, T$, out of a set of potential covariates (or lags of y_t) collected in a $T \times K$ matrix $\widetilde{W} = (\tilde{w}_1, \ldots, \tilde{w}_K)$. The selection outcome is denoted $W_\bullet = (j_T, w_1, w_2, \ldots, w_{M_\bullet})$, where j_T is a $T-$dimensional column vector of ones, $M_\bullet \leq K$, and '$\bullet$' indicates that the target set depends on some modeling strategy. Two particular strategies are encountered:

1. Specific-to-general (sg)

   (a) Initialize $W_\bullet = j_T$.

   (b) Project $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)'$ on $W_\bullet$ and obtain $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - W_\bullet \hat{q},\ \hat{q} = (W_\bullet' W_\bullet)^{-1} W_\bullet' \boldsymbol{y}$.

   (c) Estimate regressions of $\hat{\boldsymbol{\varepsilon}}$ on sets of variables $W_k = (W_\bullet,\ \tilde{w}_k),\ k = 1, \ldots, \widetilde{K}$. For each regression take an LM statistic (Godfrey 1988) measuring the explanatory content of $\tilde{w}_k$, $\lambda_k = TR_k^2$, where R_k^2 is the degree of explanation.

   (d) The particular covariate obtaining the maximum LM-statistic, λ_{k^*}, is moved from $\widetilde{W}$ to $W_\bullet$ if $\lambda_{k^*} > c_{1-\alpha}$, the $(1-\alpha)$ quantile of a $\chi^2(1)$ distribution.

   (e) Steps b) to d) are iterated until $\lambda_{k^*} \leq c_{1-\alpha}$. The search outcome is denoted W_{sg}.

2. General-to-specific (gs)

   The gs strategy proceeds from initializing $W_\bullet = \{j_T, \widetilde{W}\}_{T \times (K+1)}$, $K > (T-1)$, and subsequently eliminates the covariate $\tilde{w}_k$ from $W_\bullet$ that obtains in absolute value the smallest $t-$ratio. The gs search terminates once all variables in $W_\bullet$, denoted W_{gs}, share $\alpha\%$ significance.

3 Monte Carlo study

3.1 The data generating process

Consider the following data generating process (DGP)

$$\phi(L) y_t \ = \ (1 - \phi_1 L)(1 - \phi_4 L^4)(1 - \phi_8 L^8) y_t = \nu + u_t,\ u_t \sim iidN(0, \sigma^2). \tag{1}$$

In (1) L is short for the lag operator such that e.g. $Ly_t = y_{t-1}$. With $\nu \neq 0$ the dynamic model allows a nonzero mean, $E[y_t] = \phi^{-1}(1)\nu$. The DGP in (1) may be seen as a quarterly time series exhibiting seasonal patterns. The process is weakly stationary if all roots of $\phi(z)$ are outside the unit circle, $\phi(z) \neq 0$ for $|z| \leq 1$, implying $|\phi_1| < 1$, $|\phi_4| < 1$ and $|\phi_8| < 1$. Presuming a sufficiently large maximum lag order, p_{max}, y_t obeys the representation

$$y_t = \nu + \sum_{i=1}^{p_{max}} \underset{\sim i}{\phi} y_{t-i} + u_t.$$

Monte Carlo simulations are performed with a selection $\phi_1 = 0.8$, $\phi_4 = 0.4$, $\phi_8 = 0.2$ that implies $\underline{\phi}_1 = 0.8$, $\underline{\phi}_4 = 0.4$, $\underline{\phi}_5, = -0.32$, $\underline{\phi}_8 = 0.20$, $\underline{\phi}_9 = -0.16$, $\underline{\phi}_{12} = -0.08$, $\underline{\phi}_{13} = 0.064$, $\underline{\phi}_k = 0$, $k == 2, 3, 6, 7, 10, 11, 14, \ldots, p_{max}$. With $\sum_i \underline{\phi}_i = 0.904$ the particular parameterization generates some medium to large degree of persistence.

3.2 Evaluation criteria

Selection outcomes are evaluated with regard to ex-ante forecasting performance. Draws for y_t consist of observations $y_{-p_{max}}, \ldots, y_0, y_1, \ldots, y_{T+1}$, while T data tuples are used to determine $W_\bullet$. The last observation y_{T+1} is left for comparison of modeling implied forecast errors,

$$\hat{u}^\bullet_{T+1} = y_{T+1} - \hat{y}^\bullet_{T+1}. \tag{2}$$

Six alternative forecasting models are employed: Benchmark forecasts are derived from a 'true model' where the underlying lag structure is assumed known. Corresponding forecast errors are

$$\hat{u}^{(tm)}_{T+1} = y_{T+1} - \hat{\nu} - \sum_{i=0}^{p_{max}-1} \hat{\underline{\phi}}_i y_{T-i}, \tag{3}$$

where $\hat{\underline{\phi}}_i$ are OLS estimators obtained from a $T \times 8$ dimensional matrix W_{tm} comprising j_T and relevant lags of y_t. Owing to the consistency of $\hat{\nu}$, $\hat{\underline{\phi}}_i$, as $T \to \infty$, $E[(\hat{u}^{tm}_{T+1})^2] = \sigma^2_u$. In practical work an analyst has never access to W_{tm} but may alternatively rely on W_{sg} or W_{gs}. Implied forecast errors are denoted as $\hat{u}^{sg}_{T+1}$ or $\hat{u}^{gs}_{T+1}$, respectively. A further realistic modeling scenario is to presume that an analyst has followed both *gs* and *sg* and selects the 'final' model according to model the AIC or BIC criterion,

$$\text{AIC} = \ln(\tilde{\sigma}^2_\bullet) + (M_\bullet + 1)2/T, \;\; \text{BIC} = \ln(\tilde{\sigma}^2_\bullet) + (M_\bullet + 1)2\ln(T)/T, \;\; \tilde{\sigma}^2_\bullet = \sum_{t=1}^{T} (\hat{u}^\bullet_t)^2.$$

Modeling comparisons are based on mean squared forecast errors, $\text{MSFE}_\bullet = E[(\hat{u}^\bullet_{T+1})^2]$. A further criterion to evaluate W_{gs} against W_{sg} are empirical success probabilities

$$p_{gs} = \text{Prob}\left(|\hat{u}^{gs}_{T+1}| < |\hat{u}^{sg}_{T+1}|\right) \text{ and } p_{sg} = \text{Prob}\left(|\hat{u}^{gs}_{T+1}| > |\hat{u}^{sg}_{T+1}|\right).$$

3.3 Implementation

Alternative time series lengths $T = 30, 50, 100$ and $T = 400$ are considered to uncover how alternative modeling strategies take advantage of the consistency of the OLS estimator. Throughout time series processes are generated from zero initial conditions with disregarding the first 200 generated data points. The number of replications is 10000 throughout.

Complementary to time series modeling further simulations are performed that fit into the framework of regression models with (stationary) random explanatory variables,

$$\boldsymbol{y} = j_(T+1)\nu + X\underline{\boldsymbol{\phi}} + \boldsymbol{u}, \;\; u_t \sim iid\, N(0, \sigma^2) \tag{4}$$

	$\alpha = 0.05$						$\alpha = 0.01$		$\alpha = 0.10$	
T	*sg*	*gs*	*ac*	*bc*	p_{sg}	p_{gs}	*sg*	*gs*	*sg*	*gs*
	Autoregression, $p_{max} = 20$									
30	1.25	1.76	1.59	1.52	0.47	0.35	1.20	2.07	1.34	1.82
50	1.16	1.37	1.19	1.17	0.40	0.34	1.17	2.06	1.18	1.29
100	1.07	1.10	1.08	1.06	0.31	0.33	1.10	1.32	1.07	1.10
400	1.02	1.02	1.02	1.02	0.33	0.35	1.02	1.02	1.02	1.02
	Autoregression, $p_{max} = 14$									
30	1.12	1.48	1.18	1.17	0.41	0.32	1.13	2.14	1.14	1.31
50	1.11	1.38	1.10	1.10	0.35	0.31	1.13	2.11	1.11	1.21
100	1.06	1.07	1.05	1.05	0.27	0.28	1.09	1.33	1.05	1.07
400	1.02	1.01	1.01	1.02	0.29	0.31	1.02	1.01	1.02	1.01
	Static regression, $p_{max} = 20$									
30	1.26	1.67	1.63	1.54	0.40	0.31	1.17	1.43	1.41	1.88
50	1.18	1.22	1.21	1.20	0.20	0.18	1.14	1.25	1.23	1.26
100	1.10	1.10	1.11	1.10	0.08	0.09	1.10	1.11	1.11	1.11
400	1.02	1.02	1.02	1.02	0.02	0.02	1.02	1.02	1.02	1.02

Table 1: Forecast evaluations. Columns '*sg, gs, ac, bc*' give model specific MSFEs relative to $\text{MSFE}_{(tm)}$ obtained from the (time series) regression conditioning on the true design matrix W_{tm}. While α is the nominal level used for *gs* or *sg* empirical probabilities p_{sg} and p_{gs} indicate how often *sg* outperforms *gs* and vice versa in terms of offering smaller forecast errors.

where X is a $(T+1) \times p_{max}$ matrix drawn from a multivariate Gaussian distribution. The vector $\underline{\boldsymbol{\phi}} = \underline{\phi}_1, \ldots, \underline{\phi}_{p_{max}})'$ comprises the same parameters and zero restrictions as in the autoregressive case.

3.4 Simulation results

Table 1 documents forecasting performance for alternative model selection strategies. Most entries are relative MSFEs determined as $\text{MSFE}_\bullet/\text{MSFE}_{(tm)}$. In comparison with *sg* the *gs* approach suffers from marked inefficiency in small samples. For time series regressions with $T = 30$, $p_{max} = 20$ and $\alpha = 0.05$ *sg* implied MSFEs are about 25% in excess of the corresponding W_{tm} based quantity. The relative MSFE measure for *gs* exceeds the benchmark by 76%. The probability to get a smaller absolute forecast error from *sg* is .47 and significantly larger than $p_{gs} = .35$. Conditioning forecasts on a comparison of W_{gs} or W_{sg} implied model selection criteria mitigates the inefficiency of *gs* but remains inferior to *sg*. In case $T = 50$ *gs* is still markedly outperformed. Contrasting W_{sg} and W_{gs} by means of BIC or AIC achieves forecasting accuracy (19% and 17% excess MSFE for AIC and BIC) close to the *sg* measure of 1.16. As the sample size

further increases alternative modeling strategies perform more and more similar. Implementing iterative procedures with alternative nominal significance levels $\alpha = 0.01$ or $\alpha = 0.10$ does not affect relative outcomes. Restricting the set of potential explanatory lags by choosing $p_{max} = 14$, it turns out that all model selection strategies offer improved forecasting performance relative to $p_{max} = 20$ scenarios. Qualitatively, however, the relative performance of *sg, gs, ac, bc* is unchanged. The bottom panel of Table 1 provides corresponding results for regression modeling ($p_{max} = 20$). While the relative MSFE statistics for *sg* remain remarkably stable it appears as if *gs* performance in small samples is slightly better in static regressions as it is in autoregressions. With $T = 30$, for instance, the relative MSFE measure for the *gs* approach decreases from 1.76 to 1.67.

4 Conclusion

By means of Monte Carlo analyses specific-to-general model selection using LM tests is contrasted against a general-to-specific strategy building on OLS $t-$ratios. In small samples the former is clearly preferable to the latter in terms of modeling implied ex-ante forecasting performance, while both strategies perform similarly in extracting information from large samples. For practical purposes of variable or lag selection from large dimensional spaces, the LM based approach is featured by straightforward implementation and targeting at minimum search costs.

References

Godfrey, L.G. 1988. Misspecification Tests in Econometrics, *Cambridge University Press*, Cambridge.

Hendry, D.F., 1993. Econometrics: Alchemy or Science?, *Blackwell Publishers*, Oxford.

Hendry, D.F., Krolzig, 2001. H.-M., Automatic Econometric Model Selection with PcGets, *Timberlake Consultants Press*, London.

Lütkepohl, H. 2004. Univariate Time Series Analysis. in: Lütkepohl, H. and M. Krätzig Applied Time Series Analysis, Chapter 2, *Cambridge University Press*, Cambridge.