AutoSEARCH:
An R Package for Automated Financial Modelling*

Genaro Sucarrat†

13 March 2010

1 Introduction

This paper provides a brief documentation of the three main functions of AutoSEARCH, an R (R Development Core Team 2010) package for automated financial modelling. "SEARCH" is short for Stochastic Exponential Autoregressive Conditional Heteroscedasticity, or Stochastic Exponential ARCH, and was proposed in Sucarrat and Escribano (2010a): "Automated Model Selection in Finance: General-to-Specific Modelling of the Mean, Variance and Density". The AutoSEARCH code is based on the code developed for that project.

The are three main functions in the AutoSEARCH package. The `sm()` function is short for SEARCH model and basically estimates a SEARCH model. The second main function is `gets.mean()` and undertakes multi-path General-to-Specific (GETS) model selection of the mean specification in a SEARCH model. The third main function is `gets.var()` and undertakes multi-path GETS model selection in the variance or volatility specification of a SEARCH model.

Apart from the base packages, the only other R package AutoSEARCH depends on is the "Z's-ordered-observations" or `zoo` package (Zeileis and Grothendieck 2005, 2010), which can be downloaded from the Comprehensive R-Archive Network (CRAN, `http://CRAN.R-project.org/`). AutoSEARCH is purely written in R, mainly in version 2.6.2, but the package should work on both later versions and on earlier versions all the way back to 2.4.1, unless `zoo` version 1.5.7 or later is used. In that case R version 2.8 or later is needed. The code has been written by the undersigned, apart from the functions associated with the Exponential Power Distribution (EDP), which has been adapted from Mineo (2008).

AutoSEARCH is published under GPL version 2 or newer.

*This research was supported by a Marie Curie Intra-European Fellowship within the 6th. European Community Framework Programme, and by Banco de España.

†Department of Economics, Universidad Carlos III de Madrid (Spain). Email: gsucarra@eco.uc3m.es. Webpage: http://www.eco.uc3m.es/sucarrat/index.html.

2 The δth. power SEARCH Model

The SEARCH model belongs to the socalled ARCH class of models, a class of models that is particularly suited to model the time-varying autoregressively dependent variability often associated with financial time series. However, the versatility of the SEARCH model means it is likely to be useful in the modelling of a range of other types of data as well.

The δth. power SEARCH model consists of a mean specification, a log-variance specification and a density specification:

$$r_t = \phi_0 + \sum_{m=1}^{M} \phi_m r_{t-m} + \sum_{n=1}^{N} \eta_n x_{nt} + \epsilon_t \tag{1}$$

$$\epsilon_t = \sigma_t z_t, \quad z_t \overset{IID}{\sim} EPD(\tau), \quad \tau \geq 1 \tag{2}$$

$$\begin{aligned} \log \sigma_t^{\delta} = {} & \alpha_0 + \sum_{p=1}^{P} \alpha_p \log |\epsilon_{t-p}|^{\delta} + \sum_{a=1}^{A} \lambda_a (\log |\epsilon_{t-a}|^{\delta}) I_{\{\epsilon_{t-a}<0\}} + \omega_0 \log EWMA_{t-1} \\ & + \sum_{d=1}^{D} \omega_d y_{dt} \end{aligned} \tag{3}$$

The mean specification (1) is essentially an Mth. order autoregressive (AR) model with explanatory variables, where ϕ_0 is the mean intercept, M is the number of AR terms and N is the number of other conditioning variables that may be contemporaneous and/or lagged. Moving average (MA) terms are not included in the mean specification in order to simplify estimation and specification search. But the estimation and inference methods we employ for the log-variance specification are in general applicable subject to general conditions if MA terms are included in the mean, or if other non-linearities are included. One thing that our methods do not admit, though, is GARCH-in-mean terms.[1]

The standardised errors $\{z_t\}$ are IID and follows an Exponential Power Distribution (EPD) with shape parameter $\tau \in [1, \infty)$. Hence, when $\tau = 2$ the EPD is equal to the standard normal, when $1 < \tau < 2$ then the EPD has thicker tails than the standard normal, whereas when $\tau > 2$ then the EPD has thinner tails than the standard normal. In particular, when $\tau \to 1$ then the EPD tends to a double exponential distribution, and when $\tau \to \infty$ then the EPD tends to a uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$. An important motivation for the EPD is that, in addition to containing the normal as a special case, it allows for *both* fatter and thinner tails than the normal. The former is a common property of financial returns, whereas the latter is a real possibility in explanatory financial return modelling, since the distributional properties of $\{z_t\}$ depends on the explanatory power of the information in the mean and variance specifications (cf. Bauwens and Sucarrat 2008, and

[1] This is not necessarily a serious drawback, since proxies for financial price variability (functions of past squared returns, bid-ask spreads, function so high-low values, etc.) that can be included as regressors in the mean are readily available.

Sucarrat 2009). Another advantage of the EPD is that, under certain conditions, it ensures that the SEARCH exhibits finite moments, see Nelson (1991). By contrast, a commonly used alternative distribution for which this is not necessarily the case is the t distribution.

The logarithmic variance specification (3) is a δth. power log-ARCH model with logarithmic asymmetry terms analogous to those of Glosten et al. (1993), with the logarithm of a volatility proxy equal to an Equally Weighted Moving Average (EWMA) of past squared residuals, and with further explanatory conditioning variables y_{dt}. P is the number of log-ARCH terms, A is the number of logarithmic asymmetry terms, $EWMA_{t-1}$ is equal to $(1/T^*)\sum_{t^*=1}^{T^*}|\epsilon_{t-t^*}|^\delta$ where T^* is the length of the moving average, and D is the number of other conditioning variables that may be contemporaneous and/or lagged. If $\lambda_1 = \cdots = \lambda_A = \omega_0 = \cdots = \omega_D = 0$, then modulus greater than one for all the roots of the polynomial equation $[1 - \alpha_1 c - \cdots - \alpha_P c^P] = 0$ is a sufficient condition for stability in the variance of r_t.

Subject to suitable assumptions (stability, etc.) the parameters of the variance specifications can be estimated consistently by means of ordinary least squares (OLS) via an AR-representation. This produces a bias equal to $E(\ln z_t^2)$ (when $\delta = 2$ in the estimate of α_0, and the value of the bias depends on the distribution of z_t. However, the bias can readily be estimated by means of the residuals of the AR-representation due to the result in Sucarrat and Escribano (2010b). Furthermore, if the mean is zero or if it is estimated with sufficiently high precision in an appropriate sense, then ordinary OLS inference in the log-variance is asymptotically valid for all the parameters apart from the constant α_0.

3 The sm() function

Description

The function `sm` is used to estimate a pth. power SEARCH model.

Usage

`sm` (y, mc = NULL, ar = NULL, mx = NULL, arch = NULL, asym = NULL, log.ewma = NULL, vx = NULL, p = 2, zero.adj = 0.1, vc.adj = TRUE, varcov.mat = c("ordinary", "white"), qstat.options = NULL, tol = 1e-07, LAPACK = FALSE, verbose = TRUE, smpl = NULL)

Arguments

y	A vector, time series or zoo object that may contain NAs either at the beginning and/or at the end (the na.trim command is used to remove them), but not in the middle
mc	mc is short for "mean constant". NULL (default) does not include a constant in the mean specification, whereas any other value (say, TRUE) does
ar	integer vector that indicates the AR terms to include, for example ar=1, ar=1:4, or ar=c(2,4)
mx	vector or matrix of time series or zoo objects that contains additional regressors to be included in the mean specification
arch	integer vector that indicates the log-ARCH terms to include, for example arch=1, arch=1:3, or arch=c(3,5)
asym	integer vector that indicates the logarithmic asymmetry terms to include, for example asym=1, asym=1:4, or asym=c(2,4)
log.ewma	a list with arguments passed to the ewma function, for example log.ewma=list(length=20)
vx	vector, matrix, time series or zoo object that contains additional regressors to be included in the log-variance specification
p	numeric value (not necessarily an integer) greater than zero. The power (denoted δ) in the log-variance specification (3) above
zero.adj	quantile value used to adjust for zero-values in the residuals via the log.ep function
vc.adj	adjust the variance constant for bias (default). FALSE returns the unadjusted value

varcov.mat	sets the variance-covariance matrix used for the standard errors in the mean specification. The "white" option refers to White (1980) standard errors
qstat.options	integer vector of length two, say, c(1,1). The first value sets the order of the AR diagnostic test, whereas the second value sets the order of the ARCH diagnostic test. NULL (default) sets the vector to c(1,1)
tol	the tolerance for detecting linear dependencies in the columns of the regressors (see qr() function). Only used if LAPACK is FALSE.
LAPACK	logical. If true use LAPACK otherwise use LINPACK (see qr() function)
verbose	logical. FALSE returns less output and is faster
smpl	Either NULL (the whole sample is used for estimation) or a two-element vector of dates with the start and end dates of the sample to be used in estimation. For example, smpl=c("2001-01-01", "2009-12-31")

Value

If verbose = TRUE then a list with the following objects is returned (verbose = FALSE returns fewer objects):

call	the function call
mean.fit	zoo-object with the fitted values of the mean specification
resids	zoo-object with the residuals of the mean specification
variance.fit	zoo-object with the fitted values of the variance specification
resids.ustar	zoo-object with the residuals of the AR-representation of the log-variance specification
resids.std	zoo-object with the standardised residuals
Elogzp	the estimate of $E(\log \lvert z_t \rvert^p)$ where z_t denotes the standardised residual
Elogzstarp	the estimate of $E(\log \lvert z_t^* \rvert^p)$, see Sucarrat and Escribano (2010b). Only returned if $p \neq 2$
logEzp	the estimate of $\log E(\lvert z_t \rvert^p)$, see Sucarrat and Escribano (2010b). Only returned if $p \neq 2$
mean.results	data frame with the estimation results of the mean specification
variance.results	data frame with the estimation results of the log-variance specification

diagnostics	data frame with various diagnostic output

4 The gets.mean() function

Description

The function `gets.mean` is used to undertake GETS multi-path specification search of the mean specification of a pth. power SEARCH model.

Usage

`gets.mean`(y, mc = NULL, ar = NULL, mx = NULL, arch = NULL, asym = NULL,
log.ewma = NULL, vx = NULL, zero.adj = 0.1, vc.adj = TRUE, p = 2,
varcov.mat = c("ordinary", "white"), keep = NULL, t.pval = 0.05,
wald.pval = NULL, ar.LjungB = c(1, 0.025), arch.LjungB = c(1, 0.025),
tau = NULL, info.method = c("sc", "aic", "hq"),
info.resids = c("mean", "standardised"), include.empty = TRUE,
max.regs = 1000, tol = 1e-07, LAPACK = FALSE, verbose = TRUE,
smpl = NULL)

Arguments

The following arguments have the same structure as for the `sm()` function: y, mc, ar, mx, arch, asym, log.ewma, vx, zero.adj, vc.adj, p, varcov.mat, tol, LAPACK, verbose and smpl. Arguments specific to the `gets.mean` function are:

keep	NULL or an integer vector. If keep = NULL, then no regressors are excluded from removal. Otherwise, the regressors associated with the numbers in keep are excluded from the removal space. For example, keep=c(1) excludes the constant from removal
t.pval	numeric value between 0 and 1. The significance level used for the regressors
wald.pval	NULL or a numeric value between 0 and 1. If NULL, then no parsimonious enompassing test against the General Unrestricted Model (GUM) is undertaken. If a numeric value between 0 and 1, then a parsimonious encompassing test against the GUM is undertaken the significance level of the number between 0 and 1
ar.LjungB	NULL or a two-element vector where the first element contains the order of a Ljung and Box (1979) test for serial correlation in the standardised residuals, and where the second element contains the significance level. If NULL, then the standardised residuals are not checked for serial correlation after each removal

arch.LjungB	NULL or a two-element vector where the first element contains the order of a Ljung and Box (1979) test for serial correlation in the standardised residuals squared, and where the second element contains the significance level. If NULL, then the standardised residuals squared are not checked for serial correlation after each removal
tau	NULL or a numeric value greater than 1. If NULL, then the shape parameter of the standardised residuals is estimated for the log-likelihood used in the calculation of the information criterion. If tau is equal to a numeric value, then an $EDP(\tau)$ density is used in the computation of the log-likelihoods for the information criteria
info.method	character string, "sc", "aic" or "hq", which determines the information criterion used to select among terminal models. The abbreviations are short for the Schwarz or Bayesian information criterion (sc), the Akaike information criterion (aic) and the Hannan-Quinn information criterion
info.resids	character string, "mean" or "standardised", which sets the residuals to be used in the computation of the information criterion. If the info.resids = "mean", the default, then the residuals of the mean are used to compute the log-likelihood. If info.resids = "standardised" then standardised residuals are used
include.empty	logical equal to TRUE or FALSE. If TRUE then an empty model is included among the terminal models, if it passes the diagnostic tests. Otherwise it is not (unless one of the terminal models happens to be equal to the empty model)
max.regs	integer value, sets the maximum number of regressions along a deletion path

Value

If verbose = TRUE then a list with the following objects is returned (verbose = FALSE returns fewer objects):

resids	zoo-object with the residuals of the mean specification of the specific model
resids.std	zoo-object with the standardised residuals of the specific model
call	the function call
gum.mean	a data frame with the estimation results of the mean specification of the GUM

gum.variance	a data frame with the estimation results of the $\log \sigma_t^\delta$ specification of the GUM
gum.diagnostics	a data frame with the diagnostics of the GUM
insigs.in.gum	an integer vector referring to the insignificant regressors in the mean specification of the GUM. The length of insigs.in.gum is equal to the number of deletion paths that is searched
paths	a list of integer vectors where each vector describes the order in which the regressors are deleted along each path. A minus between the number, for example "-1", means regressor number 1 is re-included because deletion did not pass the diagnostic tests
specifications	a list of the terminal models, where each model is described by an integer vector. An integer vector equal to 0 refers to the empty model
specification.results	a data frame with the information criterion and result of the parsimonious encompassing test against the GUM associated with each terminal model
specific.mean	the specific model. A character string if empty, otherwise a data frame with the estimation results of the mean specification
specific.variance	a data frame with the estimation results of the variance specification of the specific model
specific.diagnostics	a data frame with the diagnostics of the standardised residuals of the specific model

5 The gets.var() function

Description

The function `gets.var` is used to undertake GETS multi-path specification search of the log-variance specification of a 2nd. power SEARCH model.[2]

Usage

`gets.var`(e, arch = NULL, asym = NULL, log.ewma = NULL, vx = NULL, zero.adj = 0.1, vc.adj = TRUE, keep = c(1), t.pval = 0.05, wald.pval = NULL, ar.LjungB = c(1, 0.025), arch.LjungB = c(1, 0.025), tau = NULL, tol = 1e-07, LAPACK = FALSE, info.method = c("sc", "aic", "hq"), info.resids = c("standardised", "log-sigma"), max.regs = 1000, verbose = TRUE, smpl = NULL)

Arguments

The following arguments have the same structure as for the `sm()` function: arch, asym, log.ewma, vx, zero.adj, vc.adj, tol, LAPACK, verbose and smpl, whereas the following arguments have the same structure as for the `gets.mean` function: t.pval, wald.pval, ar.LjungB, arch.LjungB, tau, info.method and max.regs. Arguments that are either specific to the `gets.var` function or which differ are:

e	A vector, time series or zoo object that contain the residuals of a mean specification. The series may contain NAs either at the beginning and/or at the end (the na.trim command is used to remove them)
keep	an integer vector that includes 1, that is, the constant of the log-variance specification is always included and excluded from removal
info.resids	character string, "standardised" or "log-sigma", which sets the residuals to be used in the computation of the information criterion. If the info.resids = "standardised", the default, then the standardised residuals are used to compute the log-likelihood. If info.resids = "log-sigma" then residuals of the AR representation of the log-variance specification are used

Value

If verbose = TRUE then a list with the following objects is returned (verbose = FALSE returns fewer objects):

[2] In future versions this will be extended to any $\delta > 0$.

ustar	zoo-object with the residuals of the AR representation of the specific model
resids.std	zoo-object with the standardised residuals of the specific model
call	the function call
gum.variance	a data frame with the estimation results of the GUM
gum.diagnostics	a data frame with the diagnostics of the GUM
keep	an integer vector of the regressors excluded from removal
insigs.in.gum	an integer vector of the insignificant regressors in the GUM. The number of regressors corresponds to the number of deletion paths that is searched
paths	a list of integer vectors where each vector describes the order in which the regressors are deleted along each path. A minus between the number, for example "-1", means regressor number 1 is re-included because deletion did not pass the diagnostic tests
specifications	a list of models that contains the terminal specifications and the GUM, all described in terms of integer vectors
specification.results	a data frame with the information criterion and result of the parsimonious encompassing test against the GUM associated with each terminal model
specific.variance	a data frame with the estimation results of the specific model
specific.diagnostics	a data frame with the diagnostics of the standardised residuals of the specific model

References

Bauwens, L. and G. Sucarrat (2008). General to Specific Modelling of Exchange Rate Volatility: A Forecast Evaluation. Forthcoming in the *International Journal of Forecasting*. UC3M Working Paper version: WP 08-18 in the Economic Series (`http://hdl.handle.net/10016/2591`).

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance*, 1779–1801.

Ljung, G. and G. Box (1979). On a Measure of Lack of Fit in Time Series Models. *Biometrika 66*, 265–270.

Mineo, A. M. (2008). *The Normalp Package*. `http://cran.r-project.org/web/packages/normalp/normalp.pdf`.

Nelson, D. B. (1991). Conditional Heteroscedasticity in Asset Returns: A New Approach. *Econometrica 51*, 485–505.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. `http://www.R-project.org`.

Sucarrat, G. (2009). Forecast Evaluation of Explanatory Models of Financial Variability. *Economics – The Open-Access, Open-Assessment E-Journal 3*. Available via: `http://www.economics-ejournal.org/economics/journalarticles/2009-8`.

Sucarrat, G. and Á. Escribano (2010a). Automated Model Selection in Finance: General-to-Specific Modelling of the Mean, Variance and Density. Available as: `http://www.eco.uc3m.es/sucarrat/research/autofim.pdf`.

Sucarrat, G. and Á. Escribano (2010b). The Power Log-GARCH Model. Available as `http://www.eco.uc3m.es/sucarrat/research/loggarch.pdf`.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica 48*, 817–838.

Zeileis, A. and G. Grothendieck (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software 14*(6), 1–27.

Zeileis, A. and G. Grothendieck (2010). *Z's ordered observations*. Vienna, Austria: R Foundation for Statistical Computing. `http://R-Forge.R-project.org/projects/zoo/`.