
ELSEVIER

Journal of Econometrics 105 (2001) 287–308

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Encompassing tests when no model is encompassing

Kenneth D. West

Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706, USA

Abstract

This paper considers regression-based tests for encompassing, when none of the models under consideration encompasses all the other models. For both in- and out-of-sample applications, I derive asymptotic distributions and propose feasible procedures to construct confidence intervals and test statistics. Procedures that are asymptotically valid under the null of encompassing (e.g., Davidson and MacKinnon, Econometrica 49 (1981) 781) can have large asymptotic and finite sample distortions. Simulations indicate that the proposed procedures can work well in samples of size typically available, though the divergence between actual and nominal confidence interval coverage sometimes is large.

JEL classification: C220; C320

Keywords: Non-nested models; V-procedure; Model selection; Forecast combination; Out of sample; Misspecification; Misspecified models

1. Introduction

It is now a truism that with sufficient data, any economic model simple enough to be analytically tractable will be rejected statistically. It is nonetheless of interest to quantify the relative explanatory powers of two or more models, even if none of the models under consideration is literally true. This will give a sense of profitable directions for future model development.

E-mail address: kdwest@facstaff.wisc.edu (K.D. West).

0304-4076/01/$ - see front matter
PII: S0304-4076(01)00078-1

Quantifying relative explanatory power can be difficult when models are nonnested, especially so when none of the models under consideration is correctly specified. A large literature has developed tests that compare in-sample fits of nonnested models. Cox's pioneering work proposed comparing likelihoods (Cox, 1961, 1962), as did Mizon and Richard (1986). Related work, on possibly misspecified models, is in Kitamura (1997). Regression based tests, involving the regression of a realization on one or more fitted values, were developed by Davidson and MacKinnon (1981). White (1994) provided a unified framework for discussing likelihood and regression-based tests, while McAleer (1995) documented the extensive use of such tests in empirical work. Finally, out-of-sample regression tests were proposed by Chong and Hendry (1986), Ericsson (1992) and West and McCracken (1998).

This paper develops asymptotic theory for regression-based encompassing tests that allow for all models under consideration to be misspecified, general classes of estimators and comparisons of out-of- as well as in-sample fits. The key result is delineation of the asymptotic variance–covariance matrix of the least squares estimator of the encompassing regression. For inference, the recommended procedure is to adjust the usual least squares variance–covariance matrix using sample analogues of the relevant asymptotic quantities—what I call the "*V-procedure*".

Section 2 of the paper uses a simple, stylized example to illustrate that construction of confidence intervals and test statistics under the incorrect null of encompassing can lead to wildly inaccurate asymptotic inference. Section 3 derives asymptotic results for least squares models, with general asymptotic results relegated to the Appendix. Section 4 presents Monte Carlo evidence. Section 5 concludes. An additional Appendix available on request presents simulation and numerical results omitted from the paper to save space.

2. Overview

The test that I consider is one in which the realization of a variable to be explained is regressed on competing in-sample fitted values or out-of-sample predictions. In out-of-sample applications, this regression is sometimes used to evaluate or combine forecasts, without reference to the word "encompassing" (see Clemens, 1989; Diebold, 1998 and especially Diebold, 1989). I nonetheless refer simply to "encompassing" tests throughout.

Suppose for simplicity that there are only two models, models 1 and 2. Write the encompassing regression as

$$y_t = \alpha_1 \hat{y}_{1t} + \alpha_2 \hat{y}_{2t} + \text{residual}. \tag{2.1}$$

Here, y_t is a scalar variable explained by models 1 and 2, $\hat{y}_{it}$ is the fitted value (or predicted value) from model i. $\hat{y}_{1t}$ and $\hat{y}_{2t}$ are constructed

from estimates of finite dimensional parameter vectors β_1 and β_2. For example, if model 1 is $y_t = X_{1t}'\beta_1 + v_{1t}$ and $\hat{\beta}_1$ is the least squares estimate, then $\hat{y}_{1t} = X_{1t}'\hat{\beta}_1$. Model 1 encompasses model 2 if $\alpha_1 = 1$, $\alpha_2 = 0$; in this case, model 2 is not helpful in explaining y_t, conditional on model 1, and model 1 gives an unbiased prediction of y_t. The symmetric condition ($\alpha_1 = 0, \alpha_2 = 1$) applies when model 2 encompasses model 1. In (2.1), a constant term, which will often be included in application, has been omitted for clarity and simplicity.

Because $\hat{y}_{1t}$ and $\hat{y}_{2t}$ depend on estimated parameters, the usual least squares estimate of the variance–covariance matrix of the estimated α's typically is not valid. (An exception to this rule is presented below.) Procedures that produce asymptotically valid in-sample tests and confidence intervals under a null of encompassing have been proposed and discussed in Davidson and MacKinnon (1981) and others.

My concern is inference about α_1 and α_2 when a null of encompassing cannot reasonably be presumed to hold. Doubt that either model is encompassing is often suggested by out of sample comparisons, or at least the initial rounds of out of sample comparisons. Such regressions often seem to suggest that none of the models are adequate. For example, in a recent study of weekly German interest rates, Ferreira (1999, p. 38) uses a set of in-sample encompassing tests to conclude that "no model ... dominates". More generally, the literature on forecast combination has repeatedly documented a failure of any single model to dominate all others (e.g., Clemens, 1989).

Of course the fundamental implication is that one needs to turn to some third (or $(n+1)^{st}$) model.[1] As a step along the way, one would like to know whether either of the two models has a lot of information about y_t. One might want to test whether one of the α_i's is zero, while not maintaining that the other α_i is unity. More generally, confidence intervals around the point estimate of the α_i's will be revealing about how well the models explain y_t.

A natural first question is whether confidence intervals constructed from conventional least squares standard errors, or from the standard errors proposed in the papers cited above, will tend to be reasonably accurate, or at least have a bias that can be characterized a priori so that rough and ready adjustment can be made. To get a feel for the answer to this question, I computed asymptotic standard errors for a simple, stylized example that affords easy calculation. This example is also used as one of the two data generating processes in the simulations.

Suppose that the data generating process is

$$y_t = x_{1t}\theta + x_{2t}(1-\theta) + u_t, \tag{2.2}$$

[1] A systematic attempt to find an encompassing model may ultimately result in a model that is the end product of extensive data mining. It is beyond the scope of this paper to consider this possibility.

$0 \leqslant \theta \leqslant 1$, where all variables are scalars and

$$(x_{1t}, x_{2t}, u_t)' \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \varrho & 0 \\ \varrho & 1 & 0 \\ 0 & 0 & \sigma_u^2 \end{pmatrix}\right). \tag{2.3}$$

Model 1 is $y_t = x_{1t}\beta_1 + v_{1t}$, model 2 is $y_t = x_{2t}\beta_2 + v_{2t}$. (In the simulations, constant terms were included in the regressions that estimated β_1 and β_2 as well as in the encompassing regression (2.1). They are omitted here because these terms do not affect asymptotic distributions.) Here, $x_{1t}\beta_1$ is the least squares projection of y_t onto x_{1t},

$$\beta_1 = (\mathrm{E}x_{1t}^2)^{-1}\mathrm{E}x_{1t}y_t = \theta + (1-\theta)\varrho, \quad v_{1t} \equiv y_t - x_{1t}\beta_1 \tag{2.4}$$

with analogous definitions for $\beta_2 = \theta\varrho + (1-\theta)$ and v_{2t}. If $\theta = 1$, model 1 encompasses model 2, and, in (2.1), $\hat{\alpha}_2$ converges in probability to zero. As well, the usual least squares standard error on $\hat{\alpha}_2$ is asymptotically valid, despite the dependence of the regressors on estimated $\hat{\beta}$'s: a very special result that holds only for $\hat{\alpha}_2$ but not $\hat{\alpha}_1$, and then only because model 1 has a scalar regressor.[2] Symmetrically, if $\theta = 0$, model 2 encompasses model 1, $\hat{\alpha}_1$ converges in probability to zero and the usual least squares standard error on $\hat{\alpha}_1$ is asymptotically valid. If $\theta \neq 0, 1$, neither model encompasses the other,

$$\begin{aligned} \hat{\alpha}_1 &\to_{\mathrm{p}} \theta/\beta_1 = \theta/[\theta + (1-\theta)\varrho], \\ \hat{\alpha}_2 &\to_{\mathrm{p}} (1-\theta)/\beta_2 = (1-\theta)/[\theta\varrho + (1-\theta)] \end{aligned} \tag{2.5}$$

and neither least squares standard error is asymptotically valid.

Here and in the simulations I consider both in-sample and out of sample fits. Suppose first that the regression (2.1) uses *in-sample* fits: one estimates β_1 and β_2 by least squares using data from 1 to T, sets $\hat{y}_{1t} = x_{1t}\hat{\beta}_1$, $\hat{y}_{2t} = x_{2t}\hat{\beta}_2$, and then estimates α_1 and α_2 by least squares using data from 1 to T. For various values of the parameters ϱ, θ and σ_u^2, I computed the asymptotic values of two estimators of the standard error on α_2 (results for α_1 are symmetric): (a) the *conventional* least squares estimate (= square root of $[\sigma_u^2 \times (2,2)$ element of the inverse of the plim of second moment matrix of regressors]), and (b) one computed in accordance with the theory presented in the next section. I used the ratio of the two to compute asymptotic coverage of nominal

[2] Davidson and MacKinnon (1981) show that when $\alpha_1 = 1$ and $\alpha_2 = 0$, the usual least squares standard error on the estimate of α_2 is asymptotically valid when one estimates $y_t - \hat{y}_{1t} = \alpha_2(\hat{y}_{2t} - \hat{y}_{1t}) + X_{1t}'a +$ residual. Here, the parameter vector "a" is not of direct interest; X_{1t} is included solely in the interest of producing a valid standard error on the estimate of α_2. But if $X_{1t} = x_{1t}$ is a scalar and $\hat{y}_{1t}$ is linear in x_{1t}, this standard error is identical to that on the estimate of α_2 in (2.1).

95 percent confidence intervals constructed using the conventional estimator. If the conventional estimator is consistent, the asymptotic coverage will be 95 percent. If the conventional estimator yields an estimate that is smaller (larger) than the valid one, asymptotic coverage will be smaller (larger) than 95 percent. For example if the asymptotic conventional estimate is about one half of the valid value, the coverage will be about 65 percent, because $\pm(0.5 \times 1.96)$ standard errors covers about 65 percent of a normal distribution.

Table 1 presents some results. As just stated, when $\theta = 1$, so that model 1 encompasses model 2, the two asymptotic values are the same: hence the "95.0" in column (5) of line 1. (Columns (6) through (10) will be explained below.) Suppose instead that $\theta \neq 1$. Begin with $\theta = 0.5$, so that the two models are equally good at explaining y_t. It may be shown analytically that compared to the appropriate value, use of conventional standard errors yields confidence intervals that are too small for when σ_u^2 is small, too large for when σ_u^2 is large. [3]

Thus, asymptotic use of the usual least squares estimate will sometimes result in spuriously narrow confidence intervals and tests that reject too frequently (small σ_u^2), sometimes result in spuriously wide confidence intervals and tests that reject too infrequently (large σ_u^2). Very small values of σ_u^2 are consistent with models of aggregate time series, in which R^2's tend to be high; large values of σ_u^2 are consistent with models of asset returns, in which R^2's tend to be low.

That this bias may be quantitatively large is suggested by the figures in column (5) in lines (2)–(7) in Table 1. The "100.0" in line (2) is a rounded figure, meaning that the coverage is $\geqslant 99.95$: the conventional standard error is much bigger (2.57 times bigger, to be exact [not reported in the Table]) than the valid one. The "64.7" in line (3) illustrates that coverage can also be far less than 95 percent. Not all specifications have such large distortions, and in some cases, least squares confidence intervals are about right (e.g., line (4)). But clearly use of OLS standard errors can lead to large distortions in either direction.

Lines (8)–(10) indicate that this holds as well when $\theta \neq 0.5$. Observe that by the symmetry in the DGP, results for inference about α_2 for given θ apply as well for inference about α_1 for $(1 - \theta)$. Hence, line (8a) tells us about inference about α_2 for $\theta = 0.8$ (i.e., when model 1 does most of the explaining about y_t) while line (8b) can be interpreted as telling us about inference about α_1 in the same regression. For this data generating process one can see in line (8) that confidence intervals constructed from the conventional standard

[3] A precise statement is that when $\theta = 0.5$, the ratio of the conventional to valid standard errors is monotonically increasing in σ_u^2, approaching a value strictly less than 1 as $\sigma_u^2 \to 0$, a value strictly greater than 1 as $\sigma_u^2 \to \infty$.

Table 1
Asymptotic coverage of nominal 95 percent confidence intervals for α_2, DGP A, least squares VCV[a]

					Coverage of nominal 95% LS conf. int. on α_2					
					In-sample	Out-of-sample				
	θ (1)	ϱ (2)	σ_u^2 (3)	R_1^2 (4)	(5)	$\pi=0.2$ (6)	$\pi=0.5$ (7)	$\pi=1$ (8)	$\pi=2$ (9)	$\pi=5$ (10)
(1)	1.0	any	any	n.a.	95.0	95.0	95.0	95.0	95.0	95.0
(2)	0.5	0.3	10.00	0.04	100.0	93.7	91.7	88.5	82.4	68.8
(3)	0.5	0.6	0.01	0.79	64.7	85.0	73.2	60.8	47.5	32.2
(4)	0.5	0.6	0.10	0.71	96.6	93.5	91.1	87.3	80.4	65.8
(5)	0.5	0.6	1.00	0.36	99.5	94.3	93.3	91.6	88.1	79.0
(6)	0.5	0.6	10.00	0.06	99.6	94.4	93.5	92.0	89.0	80.7
(7)	0.5	0.9	1.00	0.46	96.6	94.9	94.7	94.4	93.7	91.7
(8a)	0.8	0.3	0.10	0.84	89.6	92.0	87.4	80.5	70.1	52.8
(8b)	0.2			0.22	99.9	92.5	88.6	82.6	73.1	56.3
(9a)	0.8	0.6	1.00	0.45	98.1	94.8	94.6	94.1	93.2	90.4
(9b)	0.2			0.25	99.8	93.8	92.1	89.1	83.5	70.6
(10a)	0.8	0.9	1.00	0.49	95.8	95.0	94.9	94.9	94.8	94.4
(10b)	0.2			0.43	97.0	94.7	94.3	93.5	92.0	87.4

[a] *Notes*: (1) The DGP is

$$y_t = x_{1t}\theta + x_{2t}(1-\theta) + u_t, (x_{1t}, x_{2t}, u_t)' \sim \text{iid N}\left(0, \begin{pmatrix} 1 & \varrho & 0 \\ \varrho & 1 & 0 \\ 0 & 0 & \sigma_u^2 \end{pmatrix}\right).$$

If model 1 encompasses model 2, $\theta=1$; if model 2 encompasses model 1, $\theta=0$. The investigator regresses y_t on $X_{1t} \equiv (1, x_{1t})'$ and then regresses y_t on $X_{2t} \equiv (1, x_{2t})'$, obtaining coefficient estimates $\hat{\beta}_1$ and $\hat{\beta}_2$. R_1^2 is the population R^2 of the regression of y_t on X_{1t}. The final least squares regression run is the one analyzed in this table, $y_t = \alpha_0 + \alpha_1(X_{1t}'\hat{\beta}_1) + \alpha_2(X_{2t}'\hat{\beta}_2) + \text{residual}$. Here, $\hat{\alpha}_1 \to_p \theta/[\theta + (1-\theta)\varrho] \equiv \alpha_1, \hat{\alpha}_2 \to_p = (1-\theta)/[\theta\varrho + (1-\theta)] \equiv \alpha_2$. Results are invariant to omission of a constant term in any of these regressions. (2) For the indicated values of θ, ϱ and σ_u^2, columns (5)–(10) present the asymptotic coverage of nominal 95 percent confidence intervals computed using the usual least squares standard error on α_2. A value of 95.0 means that the usual least squares estimator of the standard error is consistent, a value less (greater) than 95 that this estimator yields asymptotic standard errors that are too large (small). Least squares inference can be invalid because the regressors depend on estimated $\hat{\beta}$'s. (3) Column (5) presents results when the same sample is used for obtaining the fitted values $X_{it}\hat{\beta}_i$ and the estimated $\hat{\alpha}_i$'s. Column (6)–(10) present results when an out-of-sample regression is used to estimate α_1 and α_2. The parameter "π" is the limiting ratio of the size of the out-of-sample regression (P) to the size of the samples used to estimate β_1 and $\beta_2(R)$. (4) Results for $\hat{\alpha}_1$ and given θ are identical to those for $\hat{\alpha}_2$ and $1-\theta$. For example, asymptotic in-sample coverage for $\hat{\alpha}_1$ when $\theta=0.8$, $\varrho=0.3$ and $\sigma_u^2=0.10$ is 99.9, because this is the figure in column (5) of line (8b).

error on $\hat{\alpha}_2$ are too small, those on $\hat{\alpha}_1$ too large. Subsequent lines show that sometimes both confidence intervals can be too large, sometimes both can be too small.

Turn now to an *out-of-sample* environment. I assume one step ahead forecasts for notational simplicity. As above, let T be the total amount of data available. The first R observations are used to construct $\hat{y}_{1t}$ and $\hat{y}_{2t}$; the last P observations are then used to estimate (2.1). (Other ways of dividing a data set into regression and prediction portions are discussed in the Appendix, as are multiple step ahead forecasts.) Schematically, then the sample is divided as

$$\begin{array}{lll} | \text{ ——————— } & | \text{ ——————————} | \\ 1 & R & R+P=T \end{array}. \tag{2.6}$$

I assume that realizations of right-hand side variables are used in making the prediction. (Illustration, with the AR(1) model $y_t = \beta_1 y_{t-1} + v_t$, estimated by OLS: $\hat{\beta}_1 = (\sum_{t=1}^{R} y_{t-1}^2)^{-1}(\sum_{t=1}^{R} y_{t-1} y_t)$, $\hat{y}_{1t} = y_{t-1}\hat{\beta}_1$, $t = R+1, \ldots, R+P$.)

A key parameter in the asymptotic distribution, and therefore in the simulations as well, is the limiting ratio of the size of the prediction sample to the regression sample. Call this parameter π:

$$\pi \equiv \lim_{P,\, R\to\infty} \frac{P}{R}, \quad \pi < \infty. \tag{2.7}$$

It may be shown analytically that the ratio of conventional to valid standard errors is always less than 1. Evidently, when the usual least squares estimate is used, one will obtain a spuriously narrow confidence interval, for both α_1 and α_2, at least with large samples.

The extent of the understatement is increasing in π. As $\pi \to 0$, there is no understatement; the understatement is arbitrarily large for arbitrarily large π. The natural sample analogue for π is of course P/R. In empirical work, a range of values is found, some small (e.g., $P/R \approx 0.2$ in Ericcson and Marquez, 1993), some moderate (e.g., $P/R \approx 0.4$ in Cooper, 1972), some, especially in financial applications, large (e.g., the range of values of P/R is from about 5–18 in Engle et al., 1990). Columns (6)–(10) in Table 1 show that even if one avoids the high end of this range, conventional standard errors can lead to seriously misleading inference.

3. Asymptotic theory

This section presents asymptotic results when two least squares models are compared. The Appendix spells out technical conditions relevant in this and

more general environments, including ones in which the estimation technique is GMM or maximum likelihood.[4]

Write the two models as

$$y_t = \underset{(1\times k_1)(k_1\times 1)}{X_{1t}'\beta_1} + v_{1t}, \quad y_t = \underset{(1\times k_2)(k_2\times 1)}{X_{2t}'\beta_2} + v_{2t}, \quad \beta_1 = (EX_{1t}X_{1t}')^{-1}(EX_{1t}y_t),$$

$$\beta_2 = (EX_{2t}X_{2t}')^{-1}(EX_{2t}y_t). \tag{3.1}$$

Let us allow a constant in the encompassing regression, with obvious specialization if the constant is omitted. If β_1 and β_2 were known, the encompassing regression would be

$$y_t = g_t'\alpha + u_t, \quad \underset{(3\times 1)}{\alpha} \equiv (Eg_tg_t')^{-1}(Eg_ty_t), \quad \underset{(1\times 1)}{u_t} \equiv y_t - g_t'\alpha,$$

$$\underset{(3\times 1)}{g_t} = (1, X_{1t}'\beta_1, X_{2t}'\beta_2)'. \tag{3.2}$$

Let $k = k_1 + k_2$. For simplicity I assume that the $(2k+1)\times 1$ vector $(u_t, X_{1t}'u_t, X_{2t}'u_t, X_{1t}'v_{1t}, X_{2t}'v_{2t})'$ is serially uncorrelated. This assumption consistent with many applications. An exception is in out of sample comparisons of multistep forecasts: see the Appendix for treatment of this case. In contrast to, e.g., Davidson and MacKinnon (1981), I allow for the possibility that the projection of y_t onto g_t puts nonzero values on fitted values from both models.

In practice β_1 and β_2 of course are not known. Write the corresponding least squares estimates as $\hat{\beta}_1$ and $\hat{\beta}_2$. Stack these into $(k\times 1)$ vectors $\beta = (\beta_1', \beta_2')'$ and $\hat{\beta} \equiv (\hat{\beta}_1', \hat{\beta}_2')'$. Write the fitted values as $\hat{y}_{1t} = X_{1t}'\hat{\beta}_1$ and $\hat{y}_{2t} = X_{2t}'\hat{\beta}_2$. Observe that

$$\sqrt{T}(\hat{\beta} - \beta) = \hat{B}\left(T^{-1/2}\sum_{t=1}^{T} h_t\right),$$

$$\underset{(k\times k)}{\hat{B}} = \begin{pmatrix} \left(T^{-1}\sum_{t=1}^{T} X_{1t}X_{1t}'\right)^{-1} & 0 \\ 0 & \left(T^{-1}\sum_{t=1}^{T} X_{2t}X_{2t}'\right)^{-1} \end{pmatrix},$$

$$\underset{(k\times 1)}{h_t} = \begin{pmatrix} X_{1t}v_{1t} \\ X_{2t}v_{2t} \end{pmatrix}. \tag{3.3}$$

[4] One key condition is stationarity. Restrictions on unit roots are sharper here than in the usual encompassing literature. In particular, in the standard literature, a judicious transformation allows one to use conventional inference on an encompassing test in which the variable being forecast is $I(1)$ (Fair and Shiller, 1990). This transformation is valid only under the null of encompassing, however. It does not appear, however, that there is an analogous transformation if neither model is encompassing.

Define

$$\underset{(k\times k)}{S_{hh}} \equiv \mathrm{E}h_t h_t', \ \underset{(k_1\times k_1)}{B_1} \equiv (\mathrm{E}X_{1t}X_{1t}')^{-1}, \ \underset{(k_2\times k_2)}{B_2} \equiv (\mathrm{E}X_{2t}X_{2t}')^{-1},$$

$$B_{(k\times k)} = \mathrm{diag}(B_1, B_2) = \mathrm{plim}\ \hat{B}.$$

The $(k \times k)$ asymptotic variance–covariance matrix of $\sqrt{T}(\hat{\beta} - \beta)$ is then

$$V_\beta = BS_{hh}B', \quad B' = B. \tag{3.4}$$

The sample counterpart to g_t is $\hat{g}_t = (1, X_{1t}'\hat{\beta}_1, X_{2t}'\hat{\beta}_2)'$. The least squares estimator of α is

$$\hat{\alpha} = \left(\sum_{t=1}^{T} \hat{g}_t\hat{g}_t'\right)^{-1} \sum_{t=1}^{T} \hat{g}_t y_t. \tag{3.5}$$

Substituting (3.2) into (3.5) and then using straightforward algebra yields

$$\sqrt{T}(\hat{\alpha} - \alpha) = \left(T^{-1}\sum_{t=1}^{T} \hat{g}_t\hat{g}_t'\right)^{-1} \left\{\left(T^{-1/2}\sum_{t=1}^{T} g_t u_t\right) + \left[T^{-1/2}\sum_{t=1}^{T} \hat{g}_t(g_t - \hat{g}_t)'\alpha\right] + \left[T^{-1/2}\sum_{t=1}^{T} (\hat{g}_t - g_t)u_t\right]\right\}. \tag{3.6}$$

Now, $\hat{g}_t - g_t = (0, X_{1t}'(\hat{\beta}_1 - \beta_1), X_{2t}'(\hat{\beta}_2 - \beta_2))'$. Under conditions in the Appendix,

$$T^{-1}\sum_{t=1}^{T} X_{1t}'u_t \to_{\mathrm{p}} 0, \quad T^{-1}\sum_{t=1}^{T} X_{2t}'u_t \to_{\mathrm{p}} 0, \quad \sqrt{T}(\hat{\beta}_1 - \beta_1) = \mathrm{O_p}(1),$$

$$\sqrt{T}(\hat{\beta}_2 - \beta_2) = \mathrm{O_p}(1). \tag{3.7}$$

From this it directly follows that the last term in braces on the right of (3.6) converges in probability to zero. As for the middle term,

$$T^{-1/2}\sum_{t=1}^{T} \hat{g}_t(g_t - \hat{g}_t)'\alpha = -T^{-1/2}\sum_{t=1}^{T} \hat{g}_t\alpha'(\hat{g}_t - g_t) = \tilde{F}[\sqrt{T}(\hat{\beta} - \beta)],$$

$$\tilde{F} = T^{-1}\sum_{t=1}^{T} \begin{pmatrix} 1 \\ \hat{\beta}_1'X_{1t} \\ \hat{\beta}_2'X_{2t} \end{pmatrix} (-\alpha_1 X_{1t}' - \alpha_2 X_{2t}'),$$

$$\tilde{F} \to_{\mathrm{p}} F \equiv -\begin{pmatrix} \alpha_1\mathrm{E}X_{1t}' & \alpha_2\mathrm{E}X_{2t}' \\ \alpha_1\beta_1'\mathrm{E}X_{1t}X_{1t}' & \alpha_2\beta_1'\mathrm{E}X_{1t}X_{2t}' \\ \alpha_1\beta_2'\mathrm{E}X_{2t}X_{1t}' & \alpha_2\beta_2'\mathrm{E}X_{2t}X_{2t}' \end{pmatrix}. \tag{3.8}$$

Upon combining (3.3), (3.4), (3.7) and (3.8), and using $T^{-1}\sum_{t=1}^{T}\hat{g}_t\hat{g}_t' \to_{\mathrm{p}} \mathrm{E}g_t g_t'$, we have

$$\sqrt{T}(\hat{\alpha}-\alpha) = (\mathrm{E}g_t g_t')^{-1}\left[\left(T^{-1/2}\sum_{t=1}^{T} g_t u_t\right) + FB\left(T^{-1/2}\sum_{t=1}^{T} h_t\right)\right] + \mathrm{o_p}(1). \tag{3.9}$$

Define the (3×3) matrix $S_{ff} = \mathrm{E}g_t g_t' u_t^2$ and the $3\times k$ matrix $S_{fh} = \mathrm{E}g_t h_t' u_t$. Then

$$\begin{aligned}
&\sqrt{T}(\hat{\alpha}-\alpha) \sim_{\mathrm{A}} \mathrm{N}(0,V),\\
&V = (\mathrm{E}g_t g_t')^{-1}S_{ff}(\mathrm{E}g_t g_t')^{-1} + (\mathrm{E}g_t g_t')^{-1}(FBS_{fh}' + S_{fh}B'F')\\
&\qquad (\mathrm{E}g_t g_t')^{-1} + (\mathrm{E}g_t g_t')^{-1}(FV_\beta F')(\mathrm{E}g_t g_t')^{-1}.
\end{aligned} \tag{3.10}$$

The first term is the asymptotic variance of $(\mathrm{E}g_t g_t')^{-1}(T^{-1/2}\sum_{t=1}^{T} g_t u_t)$, and is uncertainty that would be present even if β_1 and β_2 were known. The last term is the asymptotic variance of $(\mathrm{E}g_t g_t')^{-1}F[\sqrt{T}(\hat{\beta}-\beta)]$, and is attributable to uncertainty about β_1 and β_2. The middle term is the covariance between the two.

For out of sample tests, the parallel result is

$$\begin{aligned}
&\sqrt{P}(\hat{\alpha}-\alpha) \sim_{\mathrm{A}} \mathrm{N}(0,V),\\
&V = (\mathrm{E}g_t g_t')^{-1}S_{ff}(\mathrm{E}g_t g_t')^{-1} + (\mathrm{E}g_t g_t')^{-1}(\pi FV_\beta F')(\mathrm{E}g_t g_t')^{-1}.
\end{aligned} \tag{3.11}$$

The out of sample asymptotic variance is simpler because there is zero asymptotic covariance between random variables that would be present even if β_1 and β_2 were known and random variables attributable to estimation of β_1 and β_2.[5]

To further interpret (3.10) and (3.11), let V_{OLS} denote the variance–covariance matrix that would be appropriate if the β's were known rather than

[5] Recall that I am at the moment assuming that the out of sample encompassing regression is estimated using observations $R+1$ through T, while β_1 and β_2 are estimated using observations 1 through R. These samples are non-overlapping. Even in out of sample exercises, when overlapping samples are used for estimation and the encompassing regression, there is a nonzero asymptotic covariance between the two sets of random variables. See the Appendix.

estimated, $V_{\mathrm{OLS}} = (\mathrm{E}g_t g_t')^{-1} S_{ff} (\mathrm{E}g_t g_t')^{-1}$. Then (3.10) and (3.11) can be written $V = V_{\mathrm{OLS}} +$ additional terms due to estimation of β.[6] We saw in column (5) of Table 1 that in general the additional set of terms in (3.10) can raise or lower the diagonal elements of the in-sample asymptotic variance–covariance matrix. We also saw in Table 1 that for out of sample tests, the usual OLS standard errors understate the correct asymptotic ones; this is directly seen in (3.11), since $(\mathrm{E}g_t g_t')^{-1}(\pi F V_\beta F')(\mathrm{E}g_t g_t')^{-1}$ is positive semidefinite.

For inference, the obvious sample analogues can be used to estimate the additional terms in (3.10) and (3.11). The diagonal elements of the resulting estimate of V can then be used to construct confidence intervals in the usual way. I call this the "V-procedure" since it involves direct computation of the relevant variance–covariance matrix, in contrast to regression based procedures often used under the null of encompassing.

4. Monte Carlo evidence

This section uses accuracy of confidence interval coverage to get a feel for the accuracy of the asymptotic approximation developed in the previous section. Subsection 4.1 describes the data generating processes, subsection subsection 4.2 estimation and construction of the variance–covariance matrix, subsection subsection 4.3 basic results, subsection subsection 4.4 additional results.

4.1. Data generation

Two data generating processes are used. One, called "DGP A", is described in Section 2 (see Eqs. (2.2) and (2.3)). The experiments involved 36 parameter sets, where $36 = (3$ values of $\varrho) \times (3$ values of $\theta) \times (4$ values of $\sigma_u^2)$

$$\varrho = 0.3, 0.6, 0.9; \quad \theta = 0.5, 0.8, 1.0; \quad \sigma_u^2 = 0.01, 0.1, 1, 10. \tag{4.1}$$

These values were chosen for two reasons. First, they imply data whose serial- and cross-correlation properties are similar to those in Godfrey (1998) and Godfrey and Pesaran (1983) (though those authors used multivariate rather

[6] While it is not obvious (at least to me), if model 1 encompasses model 2 (i.e., $\theta = 1$, $\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 0$), and X_{1t} and X_{2t} each consist of a constant term and a scalars, the additional terms do not affect the asymptotic variance of $\hat{\alpha}_2$: the (3,3) element of $(\mathrm{E}g_t g_t')^{-1}[(FBS_{fh}' + S_{fh}B'F') + FV_\beta F'](\mathrm{E}g_t g_t')^{1}$, and of $(\mathrm{E}g_t g_t')^{-1}[\pi F V_\beta F'](\mathrm{E}g_t g_t')^{-1}$, is zero. This result is reflected in line (1) in Table 1. (N.B.: even under this special set of circumstances, the additional terms do affect the asymptotic variance of $\hat{\alpha}_1$.)

than bivariate models). Second, this range reflects certain prominent characteristics of financial and aggregate data: for financial data, competing models have low ϱ (the predictors are not very well correlated with one another) and the encompassing regression has high σ_u^2. (low R^2 in prediction of y_t); for aggregate data, competing models have high ϱ and the encompassing regression has low σ_u^2. (Of course, certain other prominent characteristics, such as serial correlation or conditional heteroskedasticity, are not captured by this process. Since these complications probably degrade the quality of the asymptotic approximation for given sample size, the results here may be unduly supportive.)

The second data generating process, called "DGP B", involved comparison of models linear in the level and in the log of an explanatory variable. The motivation was twofold. First, encompassing tests are used in practice to discriminate between log and semilog specifications (e.g., Stumborg, 1999). Second, simulation evidence on encompassing tests indicates that the tests sometimes perform poorly when non-normal data are used (e.g., Godfrey, 1998). So evaluation of the V-procedure for a non-symmetric (specifically, lognormal) variable seemed advisable.

DGP B was

$$y_t = \theta x_t + (1-\theta)\ln(x_t) + u_t,\ (\ln(x_t), u_t)' \sim \text{iid N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}\right). \tag{4.2}$$

The two competing models are

$$y_t = \beta_{01} + \beta_{11} x_t + v_{1t} \equiv X_{1t}'\beta_1 + v_{1t}, \tag{4.3a}$$

$$y_t = \beta_{02} + \beta_{12} \ln(x_t) + v_{2t} \equiv X_{2t}'\beta_2 + v_{2t}. \tag{4.3b}$$

The experiments with this DGP involved 20 parameter sets, (5 values of θ) $\times$ (4 values of σ_u^2)

$$\theta = 0, 0.2, 0.5, 0.8, 1.0; \sigma_u^2 = 0.01, 0.1, 1, 10. \tag{4.4}$$

There is no variation in ϱ because the correlation between the two regressors is not a free parameter; in all specifications considered, $\text{corr}(x_t, \ln(x_t)) \approx 0.76$. In addition, the results for α_1 and α_2 are no longer symmetric, so results for both are presented. Finally, to save space, I report results only for $\sigma_u^2 = 0.1$, reporting complete results in the additional Appendix.

For each DGP and parameter set, I generated 5000 samples of size 500. Only the first $T = 100$ or first $T = 250$ were used in the in-sample experiments. For the out-of-sample work, there were 6 different sets of regression and prediction sample sizes: $R = 100, P = 50$; $R = 100, P = 100$; $R = 100, P = 200$; $R = 250, P = 50$; $R = 250, P = 125$; $R = 250, P = 250$. I report only results for $R = 100$. Results for $R = 250$ were similar and are reported in the additional Appendix. I also conducted some out of sample simulations using

what the Appendix calls the "recursive" scheme; I report these in the additional Appendix but not here since results are similar to those reported in the tables below.

4.2. Estimation

For the in-sample test, I used each of these two samples ($T=100$ and $T=250$) as follows. (1) Obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ by least squares regressions of y_t on X_{1t} and $X_{2t}, t=1,\ldots,T$. (2) Estimate α_1 and α_2 in a least squares regression of y_t on a constant, $X_{1t}'\hat{\beta}_1$ and $X_{2t}'\hat{\beta}_2$. (The transpose "′" is needed even for DGP A, since constant terms were included in all regressions: for DGP A, $X_{it} \equiv (1, x_{it})'$.) (3) Compute two different variance–covariance matrices. The first is the usual heteroskedasticity consistent covariance matrix for least squares. The second is an estimate of V defined in (3.10), constructed as described below. (4)(a) DGP A: use the estimated variance–covariance matrices to construct 95 percent confidence intervals around $\hat{\alpha}_2$. Report the percentage of confidence intervals that actually include $\alpha_2 \equiv (1-\theta)/[\theta\varrho \pm (1-\theta)]$. (b) DGP B: use the estimated variance–covariance matrices to construct 95 percent confidence intervals around $\hat{\alpha}_1$ and $\hat{\alpha}_2$. Report the percentage of confidence intervals that actually include the population values of α_1 and α_2, which happen to be $\alpha_1 = \theta/\{\theta+[(1-\theta)\sqrt{e}/(e^2-e)]\}$, $\alpha_2 = (1-\theta)/[(\theta\sqrt{e})+(1-\theta)]$. (For both DGPs, the additional Appendix reports results for 90 percent confidence intervals, which were similar.)

Inference was done with heteroskedasticity consistent covariance matrices, even though there is no heteroskedasticity in the disturbance in the encompassing regression u_t. To spell out the details, some notation has to be defined. In the encompassing regression $y_t = \alpha_0 + \alpha_1(X_{1t}'\hat{\beta}_1) + \alpha_2(X_{2t}'\hat{\beta}_2)$+residual, define the vector of right-hand side variables, least squares coefficient estimates and scalar residual as

$$\underset{3\times 1}{\hat{g}_t} \equiv (1, X_{1t}'\hat{\beta}_1, X_{2t}'\hat{\beta}_2)', \quad \underset{3\times 1}{\hat{\alpha}} \equiv (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)', \quad \underset{(1\times 1)}{\hat{u}_t} = y_t - \hat{g}_t'\hat{\alpha}. \tag{4.5}$$

Also define

$$\underset{(4\times 1)}{\hat{h}_t} = (X_{1t}'\hat{v}_{1t}, X_{2t}'\hat{v}_{2t})', \quad \underset{(4\times 1)}{\frac{\partial \hat{u}_t}{\partial \beta}} = -(\hat{\alpha}_1 X_{1t}'\hat{\alpha}_2 X_{2t}')', \tag{4.6}$$

In (4.6), $\hat{v}_{1t}$ and $\hat{v}_{2t}$ are least squares residuals and $\hat{h}_t$ is the sample cross product of right hand side variables and residuals in the regressions used to estimate β_1 and β_2.

For in-sample confidence intervals, define the (2×2) matrices $\hat{B}_1 = (T^{-1}\sum_{t=1}^{-T} X_{1t}X_{1t}')^{-1}$, $\hat{B}_2 = (T^{-1}\sum_{t=1}^{T} X_{2t}X_{2t}')^{-1}$. The sample analogues of the population quantities that figure into V were estimated as follows:

$Eg_t g_t'$: $T^{-1}\ \sum_{t=1}^{T} \hat{g}_t \hat{g}_t'$; S_{ff}: $T^{-1} \sum_{t=1}^{T}\ \hat{g}_t \hat{g}_t' \hat{u}_t^2$; S_{fh}: $T^{-1} \sum_{t=1}^{T} (\hat{g}_t \hat{u}_t) \hat{h}_t'$; F: T^{-1} $\sum_{t=1}^{T} [\hat{g}_t (\partial \hat{u}_t / \partial \beta)']$; B: $\text{diag}(\hat{B}_1, \hat{B}_2)$; V_β: $\hat{B}(T^{-1} \sum_{t=1}^{T} \hat{h}_t \hat{h}_t')^{-1} \hat{B}$. For out of sample confidence intervals, B and V_β were estimated using data from 1 to R; $Eg_t g_t'$, S_{ff}, F and S_{fh} were estimated with data running from $R+1$ to $R+P$.

For certain experiments I also report confidence intervals constructed from the usual heteroskedasticity consistent least squares estimator. This was constructed as

$$V_{\text{OLS}}: \left(T^{-1} \sum_{t=1}^{T} \hat{g}_t \hat{g}_t' \right)^{-1} \left(T^{-1} \sum_{t=1}^{T} \hat{g}_t \hat{g}_t' \hat{u}_t^2 \right) \left(T^{-1} \sum_{t=1}^{T} \hat{g}_t \hat{g}_t' \right)^{-1}. \qquad (4.7)$$

When the null of encompassing holds, inference using V_{OLS} is asymptotically valid, and is consistent with Davidson and MacKinnon (1981).[7]

4.3. Simulation results

Results for DGP A are reported in Table 2. In-sample results are presented in columns (4) and (5). For $T=100$, a couple of parameterizations lead to results that are troubling, for example the coverage rate of 92.0 reported in line (2), column (4). This is consistent with the still worse results reported for sizes of $T=40$ and 60 by Godfrey (1998). But for $T=250$ all but one of the reported results are between 94 and 96. The out of sample tests reported in lines (6) through (8) are similar. All involve regression sample size $R=100$, and all have some parameterizations with poor coverage. Out of sample results for $R=250$ (reported in the additional Appendix) are comparable to in sample results for $T=250$. But even for $R=100$, on balance the figures are tolerably close to 95.

The V-procedure does not fare as well in the second experiment. Representative results are given in Table 3. Separate results are given for α_1 and α_2 because the results are no longer symmetric. In Panel B, columns (3)–(7), the figures for α_2 in range from 89.0 to 94.5, somewhat less satisfying than previously. The news about α_1 in Panel A, columns (3)–(7), is still worse, with figures as low as 76.0 (line (5), $\theta=0.5$).

It may be little consolation, but inference using the conventional heteroskedasticity consistent least squares covariance matrix was even more awry. Begin with DGP A, for which the V-procedure worked well. Panel A in Table 4 has representative results. Many of the figures are far from 95.

[7] To illustrate with DGP A, when $\theta=1$: in the spirit of Davidson and MacKinnon's J-test, one could estimate $y_t = \alpha_0 + \delta x_{1t} + \alpha_2 (X_{2t}' \hat{\beta}_2) + \text{residual}$, and test $H_0: \delta = 0$. This test is identical to the results I report for least squares inference about α_1 in $y_t = \alpha_0 + \alpha_1 (X_{1t}' \hat{\beta}_1) + \alpha_2 (X_{2t}' \hat{\beta}_2) + \text{residual}$, with $X_{1t} \equiv (1, x_{1t}')'$. This is not quite the J-test, and inclusion of the constant term may degrade finite sample performance.

Table 2
Actual coverage of nominal 95 percent confidence intervals for α_2, DGP A, V-procedure[a]

				Coverage of 95% confidence interval on α_2				
				In-sample		Out-of-sample, $R=100$		
	θ (1)	ϱ (2)	σ_u^2 (3)	$T=100$ (4)	$T=250$ (5)	$P/R=0.5$ (6)	$P/R=1$ (7)	$P/R=2$ (8)
(1)	1.0	0.6	1.00	94.2	95.4	94.7	96.5	96.9
(2)	0.5	0.3	10.00	91.7	94.0	95.3	93.9	92.1
(3)	0.5	0.6	0.01	93.7	94.3	93.7	94.0	94.0
(4)	0.5	0.6	0.10	93.6	94.2	93.4	94.5	94.4
(5)	0.5	0.6	1.00	93.7	95.0	93.6	94.8	94.9
(6)	0.5	0.6	10.00	95.7	96.2	96.1	95.7	94.3
(7)	0.5	0.9	1.00	94.5	94.8	92.6	94.8	95.2
(8a)	0.8	0.3	0.10	95.7	95.7	94.0	95.7	94.2
(8b)	0.2	0.3	0.10	91.4	94.1	93.3	94.5	94.4
(9a)	0.8	0.6	1.00	94.5	95.1	93.8	95.4	95.5
(9b)	0.2	0.6	1.00	94.2	95.3	93.4	94.8	95.0
(10a)	0.8	0.9	1.00	94.4	95.0	92.6	94.9	95.0
(10b)	0.2	0.9	1.00	94.5	94.8	92.3	93.9	93.9

[a]*Notes*: (1) The data generating process is described in Table 1. In columns (4) and (5), T is the sample size. In columns (6)–(8), $R=100$ is the size of the sample used to obtain the least squares estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ (defined in note 1 to Table 1), while P is the size of the sample used to obtain the least squares estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ (again defined in note 1 to Table 1). All results are based on 5000 repetitions. (2) The V-procedure uses sample analogues to estimate the quantities in asymptotic variance–covariance matrices presented in Eqs. (3.10) and (3.11), and then uses the diagonal elements of these matrices to construct confidence intervals in the usual way. See Section 4.2 for details. This procedure will yield asymptotic coverage rates of 95.0. (3) Results for $\hat{\alpha}_1$ are symmetric to those for $\hat{\alpha}_2$, as explained in the notes to Table 1.

For example, we see in line (2) that for $T=250$, the least squares confidence interval has coverage of 63.8 percent; in Table 2, the comparable figure using the V-procedure is 94.3. Upon comparing Tables 1 and 4A, we see that figures such as 63.8 reflect the asymptotic theory. This theory does quite a good job of predicting which intervals will be too short and which will be too long: for both in- and out-of-sample exercises, the asymptotic theory and

Table 3
Coverage of nominal 95 percent confidence intervals for α_1 and α_2, DGP B, V-procedure[a]

				Coverage of 95% confidence interval on α_2				
				In-sample		Out-of-sample, $R=100$		
		θ (1)	σ_u^2 (2)	$T=100$ (3)	$T=250$ (4)	$P/R=0.5$ (5)	$P/R=1$ (6)	$P/R=2$ (7)
(A) α_1	(1)	1.0	0.10	93.5	94.8	86.4	88.9	90.4
	(2)	0.8	0.10	85.6	86.6	82.5	82.6	81.8
	(3)	0.5	0.10	76.0	77.2	83.4	84.0	83.9
	(4)	0.2	0.10	77.5	78.0	76.4	71.9	67.2
	(5)	0.0	0.10	88.9	90.5	87.2	89.0	90.7
(B) α_2	(1)	1.0	0.10	93.3	94.2	91.8	94.1	94.5
	(2)	0.8	0.10	90.7	92.6	90.6	91.2	92.1
	(3)	0.5	0.10	89.9	91.7	91.2	92.5	92.7
	(4)	0.2	0.10	90.4	92.3	91.2	92.0	92.1
	(5)	0.0	0.10	89.0	89.7	91.6	93.3	94.0

[a]*Notes*: (1) The DGP is

$$y_t = x_t\theta + \ln(x_t)(1-\theta) + u_t,\ (\ln(x_t), u_t)' \sim \text{iid N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}\right),\ \sigma_u^2 = 0.1.$$

If model 1 encompasses model 2, $\theta=1$; if model 2 encompasses model 1, $\theta=0$. The investigator first regresses y_t on $X_{1t} \equiv (1, x_t)'$ and then on $X_{2t} \equiv (1, \ln(x_t))'$, obtaining 2×1 coefficient vectors $\hat{\beta}_1$ and $\hat{\beta}_2$. The final least squares regression run is the one whose results are analyzed in this table, $y_t = \alpha_0 + \alpha_1(X_{1t}'\hat{\beta}_1) + \alpha_2(X_{2t}'\hat{\beta}_2) + \text{residual}$. Here, $\hat{\alpha}_2 \to_p \theta/\{\theta + [(1-\theta)\sqrt{e}/(e^2-e)]\} \equiv \alpha_1, \hat{\alpha}_2 \to_p (1-\theta)/[(\theta\sqrt{e}) + (1-\theta)] \equiv \alpha_2$. (2) See notes to Table 2.

simulations match perfectly on whether coverage is less than or greater than 95 percent, and this holds for all the specifications in Table 1 and not just the subset reported in Table 4A.

Panel B in Table 4 indicates that conventional procedures also fared quite poorly for DGP B, even more poorly than did the V-procedure. For example, in the specification that the V-procedure performed worst ($\theta=0.5$), with in-sample coverage of 89.9 percent for $T=100$, least squares coverage was 79.3 (see panel B, line (3), column (4)). The corresponding asymptotic figures in panel C indicate that poor coverage is to be expected for least squares-indeed, for big enough samples the 79.3 figure will fall to 74.2 (panel C, line (3), column (4)).

4.4. Additional simulation results

To get a sense for rapidly increases in-sample size lead to improvements in the accuracy of the asymptotic approximation, I picked the worst performing

Table 4
Coverage of nominal 95 percent confidence intervals for α_2, least squares VCV[a]

					Coverage of 95% confidence interval on α_2				
					In-sample		Out-of-sample, $R=100$		
		θ (1)	ϱ (2)	σ_u^2 (3)	$T=100$ (4)	$T=250$ (5)	$P/R=0.5$ (6)	$P/R=1$ (7)	$P/R=2$ (8)
(A)Actual coverage, DGP A	(1)	1.0	0.6	1.00	94.0	94.6	92.1	94.3	94.8
	(2)	0.5	0.6	0.01	63.0	63.8	70.1	59.3	46.9
	(3)	0.5	0.6	0.10	95.6	96.3	87.9	85.4	80.6
	(4)	0.5	0.6	1.00	99.1	99.3	90.2	90.3	87.5
	(5)	0.5	0.6	10.00	99.4	99.6	90.5	90.9	88.2
(B) Actual coverage, DGP B	(1)	1.0	n.a.	0.10	92.3	93.6	90.3	92.6	93.4
	(2)	0.8	n.a.	0.10	86.5	87.7	87.1	86.2	82.4
	(3)	0.5	n.a.	0.10	79.3	77.9	82.0	76.0	66.5
	(4)	0.2	n.a.	0.10	85.8	86.2	84.4	80.4	72.6
	(5)	0.0	n.a.	0.10	96.9	98.0	88.4	87.7	83.6
(C) Asymptotic coverage, DGP B	(1)	1.0	n.a.	0.10	95.0	95.0	95.0	95.0	95.0
	(2)	0.8	n.a.	0.10	87.8	87.8	90.8	86.6	79.2
	(3)	0.5	n.a.	0.10	74.2	74.2	81.7	71.8	59.1
	(4)	0.2	n.a.	0.10	84.6	84.6	86.2	78.6	67.5
	(5)	0.0	n.a.	0.10	99.0	99.0	92.5	90.0	85.2

[a]*Notes*: (1) The data generating processes are described in Tables 1 and 2. (2) In panels A and B, confidence intervals were constructed from the usual heteroskedasticity consistent least squares variance–covariance matrix. For panel A, this will give asymptotic coverage rates given in Table 1, lines (1), (3)–(5) and (6).

specification from Tables 2 and 3, $\theta=0.5$, DGP B, and experimented with in-sample inference with larger sample sizes. The results for $T=1000, 2500$ and 10,000 are given in panel A of Table 5, with results for $T=100$ and 250 repeated for convenience. Naturally, the asymptotic approximation works better with larger samples. For example, for $T=2500$, inference about α_2, using either the proposed or the usual least squares inference, works pretty much in accord with the asymptotic theory. (For least squares this follows since the figure of 75.8 for $T=2500$ is quite near the asymptotic figure of 74.2 reported in panel C of Table 4.) But while inference about α_1 is better captured by the approximation for larger T, even for $T=10,000$ there are notable discrepancies, for the V-procedure (actual $=88.9$, asymptotic $=95.0$) or least squares (actual $=38.9$, asymptotic $=29.8$ [not reported in a Table]).

I therefore briefly consider bootstrapping the V-procedure. I constructed confidence intervals from symmetric two tailed t-tests, with 500 bootstrap repetitions per sample, again with 5000 samples. Each bootstrap repetition involved resampling to generate new estimates of the β's as well as of the α's. Details on the procedure are given in the additional Appendix.

Table 5
Additional simulation results on 95 percent confidence interval coverage, in-sample tests[a]

(A) DGP B, actual coverage, large sample sizes

		$T=100$		$T=250$		$T=1000$		$T=2500$		$T=10,000$	
	θ	V	OLS	V	OLS	V	OLS	V	OLS	V	OLS
(1) α_1	0.5	76.0	61.2	77.2	54.0	81.3	47.1	85.0	43.0	88.4	38.9
(2) α_2	0.5	89.9	79.3	91.7	77.9	93.2	76.4	94.2	75.8	94.4	74.9

(B) DGP A, actual coverage for α_2

	$T=100$			$T=250$		
θ	BS-V	V	OLS	BS-V	V	OLS
1.0	94.8	94.2	94.0	95.3	95.4	94.6
0.5	94.9	93.7	99.1	95.2	95.0	99.3

(C) DGP B, actual coverage for α_1 and α_2

		$T=100$			$T=250$		
	θ	BS-V	V	OLS	BS-V	V	OLS
(1)α_1	1.0	95.0	93.5	92.9	95.4	94.8	95.5
	0.5	84.0	76.0	61.2	83.6	77.2	54.0
(2)α_2	1.0	94.9	93.3	92.3	94.7	94.2	93.6
	0.5	92.9	89.9	79.3	93.4	91.7	77.9

[a]*Notes*: (1) "BS-V" denotes confidence intervals constructed by bootstrapping the V-procedure, via symmetric two-tailed t-statistics; "V" denotes the procedure proposed in this paper; "OLS" denotes confidence intervals constructed from a heteroskedasticity consistent least squares covariance matrix. The results for "V" and "OLS" are repeated from Tables 3 and 4. (2) See notes to Tables 1 and 3 for descriptions of the data generating processes. In panel B, $\varrho=0.6$ and $\sigma_u^2=1.0$; in panels A and C, $\sigma_u^2=0.1$. All results are based on 5000 repetitions. For BS, there were 500 bootstrap repetitions for each of the 5000 samples.

I report representative results in panels B and C of Table 5. The figures for the V-procedure and for least squares repeat those given in Tables 2–4, for convenience. The $\theta=1$ lines in panels B and C indicate that all three procedures (bootstrap of V-procedure, V-procedure, least squares) work roughly comparably under the null of encompassing, with bootstrapping having an edge. For example, for DGP B, $T=100$ panel C indicates that bootstrapping happened to be spot on, with actual coverage of 95.0 percent; the coverage of the other procedures ranged from 92.3 to 93.5. For $\theta=0.5$, least squares inference is asymptotically invalid. Upon comparing the bootstrap and the regular versions of the V-procedure, we see that the bootstrapped version performs better, markedly so for DGP B. We see in panel C, line 1 that bootstrap coverage when $\theta=0.5$ is around 84 percent. That is far from the

ideal of 95 percent but still is a distinct improvement over the figures of 76.0 and 77.2 for the V-procedure.

5. Conclusions

Regression-based tests for encompassing were proposed and evaluated. The tests allow for the possibility that none of the models under consideration encompass the others. Simulations indicate that V-procedure can work well, though there sometimes are notable distortions. Even when there are notable distortions, the V-procedure usually works better than does a conventional procedure that is asymptotically valid only when the null of encompassing holds. A priority for future research is developing refined procedures that provide a more accurate guide to performance in small samples. Limited simulation evidence suggests that bootstrapping may deliver such procedures.

Acknowledgements

I thank the National Science Foundation for financial support, Frank Diebold and an anonymous referee for helpful comments, and Stanislav Anatolyev, Gabriel Di Bella and Mukunda Sharma for research assistance.

Appendix

I begin by extending the environment described in the text in three ways, and then present formal conditions that lead to a general result that includes Eqs. (3.10) and (3.11) as special cases. First, out of sample tests are sometimes executed allowing multiperiod predictions. Let us therefore allow for a prediction horizon $\tau \geqslant 1$ periods ahead (the text assumed $\tau = 1$). If the null of encompassing holds, $u_t \sim \text{MA}(\tau - 1)$. Let the total sample size be $T = R + P + \tau - 1$.

Second, in out-of-sample studies, let us allow two more ways of splitting a sample into regression and prediction portions. The *rolling* scheme uses the last "R" observations to estimate the two models. It first uses data from 1 to R to estimate the models and predict $y_{R+\tau-1}$, then uses data from 2 to $R+1$ to estimate the models and predict $y_{R+\tau}, \ldots$ and finally uses data from P to $R+P-1$ to estimate the models and predict $y_{P+R+\tau-1}$. The *recursive* scheme uses a growing sample size to estimate the two models, first using data from 1 to R, then from 1 to $R+1, \ldots$, and finally from 1 to $R+P-1$. As a matter of terminology, the division described in the text is called the *fixed* scheme.

Third, let us allow encompassing tests that involve more than two models. Write the $(n+1)\times 1$ vector of right hand side variables as $\hat{g}_t = (1, \hat{y}_{1t}, \ldots, \hat{y}_{nt})'$.

To state formal assumptions, it will be helpful to denote the population parameter vector, obtained by stacking the parameters from each of the n models, as β^* rather than β. Additional notation: $u_{t\beta}(\beta^*)$ is the $(1\times k)$ matrix $\partial u_t(\beta^*)/\partial\beta$; $g_{t\beta}(\beta^*)$ is the $(n\times k)$ matrix $\partial g_t(\beta^*)/\partial\beta$; for any matrix $A=[a_{ij}]$, let $|A|\equiv \max_{i,j}|a_{ij}|$. The assumptions in West (1996) and West and McCracken (1998, p. 822) are sufficient for my purpose:

Assumption (*):. (a)(i)In some neighborhood N around β^*, and with probability 1, $u_t(\beta)$ and $g_t(\beta)$ are measurable and twice continuously differentiable; (ii) $Eu_t(\beta^*)g_t(\beta^*)=0$; (iii) $Eu_t(\beta^*)u_{t\beta}(\beta^*)=0$; (iv) $Eu_t(\beta^*)g_{t\beta}(\beta^*)=0$; (v) $Eg_t(\beta^*)g_t(\beta^*)'$ has rank $n+1$.

(b)(i)The estimate $\hat\beta_t$ satisfies $\hat\beta_t-\beta^*=\hat B(t)H(t)$, where $\hat B(t)$ is $(k\times q)$ and $H(t)$ is $(q\times 1)$, with (a)$\hat B(t)\to_{\text{a.s.}} B$, B a matrix of rank k; (ii)$H(t)=T^{-1}\sum_{s=1}^{T} h_s(\beta^*)$ (in sample), $H(t)=R^{-1}\sum_{s=1}^{R} h_s(\beta^*)$ (fixed), $H(t)=t^{-1}\sum_{s=1}^{t} h_s(\beta^*)$ (recursive), or, $H(t)=R^{-1}\sum_{s=t-R+1}^{t} h_s(\beta^*)$ (rolling) for a $(q\times 1)$ orthogonality condition $h_s(\beta^*)$; (iii) $Eh_s(\beta^*)=0$; (iv) in the neighborhood N of assumption 1, h_t is measurable and continuously differentiable.

(c) In the neighborhood N of Assumption 1, there is a constant $D<\infty$ such that for all t, $\sup_{\beta\in N}|\partial^2 u_t(\beta)/\partial\beta\partial\beta'|<m_t$ for a measurable m_t for which $Em_t^4<D$. The same holds when u_t is replaced by an arbitrary element of g_t.

(d) Let $w_t\equiv(u_{t\beta}(\beta^*)',\ \mathrm{vec}(g_{t\beta}(\beta^*))',\ u_t(\beta^*),\ g_t(\beta^*)',\ h_t(\beta^*)')'$. (i) For some $d>1$, $\sup_t E||w_t||^{8d}<\infty$, where $||\cdot||$ denotes Euclidean norm. (ii) w_t is strong mixing, with mixing coefficients of size $-3d/(d-1)$. (iii) w_t is fourth order stationary. (iv) $\sum_{j=-\infty}^{\infty} E[g_t(\beta^*)g_{t-j}(\beta^*)'u_t(\beta^*)u_{t-j}(\beta^*)]$ is positive definite.

(e) For out-of-sample tests, $R,P\to\infty$ as $T\to\infty$, and $\lim_{T\to\infty}P/R=\pi$, (i) $0\leqslant\pi\leqslant\infty$ for recursive, (ii) $0\leqslant\pi<\infty$ for rolling and fixed.

A word on the assumptions. Assumption (a) essentially says that u_t is orthogonal to the predictors from all the models. For example, in the linear models of Section 3, $u_{t\beta}=(\partial/\partial\beta)[y_t-\alpha_1(X_{1t}'\beta_1)-\alpha_2(X_{2t}'\beta_2)]=(-\alpha_1X_{1t}'-\alpha_2X_{2t}')$, so $Eu_t(\beta^*)u_{t\beta}(\beta^*)=0$ means $Eu_tX_{it}'=0$. As well, the rank condition on $Eg_t(\beta^*)g_t(\beta^*)'$ rules out nested models such as $y_t=X_{1t}\beta_1+v_t$ vs. $y_t=X_{1t}\beta_1+Z_t\delta+v_t$ with population $\delta=0$.

Assumption (b): the underlying assumption is that the estimate from the i'th model can be written $\hat\beta_{it}-\beta_i=\hat B_i(t)H_i(t)$ for $\hat B_i(t)$ and $H_i(t)$ illustrated below. $\hat B(t)$ is a block-diagonal matrix with diagonal blocks $\hat B_i(t)$; $H(t)$ is obtained by stacking $H_1(t),\ldots,H_n(t)$. As is evident from the definitions of $H(t)$, the "t" index is not necessary for $H(t)$ for in-sample applications (i.e., for given sample size T, $H(1)=\cdots=H(T)$), nor for out of sample applications using the fixed scheme; the same applies to $\hat B_i(t)$ and consequently $\hat B_{it}$. I use the index nonetheless because it is necessary for the recursive and rolling schemes. See West and McCracken (1998) for examples.

For maximum likelihood, h_{it} is the score, evaluated at the population parameter vector β_i, and $q_i = k_i$. For GMM, h_{it} is the set of moment conditions used to identify β_i (e.g., the Kronecker product of the vector of predetermined variables and the vector of structural disturbances, if the estimator is 3SLS), and $q_i \geqslant k_i$. $\hat{B}_i(t)$ is a $(k_i \times q_i)$ matrix of rank k_i that selects a linear combination of orthogonality conditions. For maximum likelihood, $\hat{B}_i(t)$ is the inverse of the Hessian, evaluated on the line between β_{it} and $\hat{B}_i$; for GMM in overidentified systems, $\hat{B}_i(t)$ depends on the weighting matrix used (see Hansen, 1982). B is the large sample counterpart of $\hat{B}(t)$. See Section 3 for concrete illustration for least squares models.

Assumptions (c)–(e) are technical conditions whose main practical import is to rule out models with unit autoregressive roots.

Define $f_t(\beta^*) = g_t(\beta^*)u_t(\beta^*)$ $[(n+1)\times 1]$, $F = \mathrm{E}[g_t(\beta^*)u_{t\beta}(\beta^*)]$ $[(n+1)\times k]$, $S_{ff} = \sum_{j=-\infty}^{\infty} \mathrm{E} f_t(\beta^*) f_{t-j}(\beta^*)'$ $[(n+1)\times(n+1)]$, $S_{fh} = \sum_{j=-\infty}^{\infty} \mathrm{E} f_t(\beta^*) h_{t-j}(\beta^*)'$ $[(n+1)\times q]$, $S_{hh} = \sum_{j=-\infty}^{\infty} \mathrm{E} h_t(\beta^*) h_{t-j}(\beta^*)'$ $[q\times q]$. In out-of-sample evaluation of τ step ahead forecasts, $f_t \sim \mathrm{MA}(\tau-1)$; in most applications, f_t (and h_t) are serially uncorrelated, so that $S_{ff} = E f_t(\beta^*) f_t(\beta^*)'$, $S_{fh} = \mathrm{E} f_t(\beta^*) h_t(\beta^*)'$ and $S_{hh} = \mathrm{E} h_t(\beta^*) h_t(\beta^*)'$: see the least squares example in Section 3. Also define the scalars λ_{fh} and λ_{hh} as follows. For in-sample tests, $\lambda_{fh} = 1$, $\lambda_{hh} = 1$. For out-of-sample tests: recursive, $\lambda_{fh} = 1 - \pi^{-1}\ln(1+\pi)$, $\lambda_{hh} = 2[1-\pi^{-1}\ln(1+\pi)]$; fixed: $\lambda_{fh} = 0$, $\lambda_{hh} = \pi$; rolling, $\pi \leqslant 1$, $\lambda_{fh} = \pi/2$, $\lambda_{hh} = \pi - \pi^2/3$, rolling, $\pi > 1$, $\lambda_{fh} = 1 - 1/2\pi$, $\lambda_{hh} = 1 - 1/3\pi$.

Theorem. Under Assumption $(*)$, $\sqrt{T}(\hat{\alpha}-\alpha)$ *(in-sample) and* $\sqrt{P}(\hat{\alpha}-\alpha)$*(out-of-sample) are asymptotically normal with variance–covariance matrix*

$$V = (\mathrm{E} g_t g_t')^{-1} S_{ff} (\mathrm{E} g_t g_t')^{-1} + (\mathrm{E} g_t g_t')^{-1} [\lambda_{fh}(FBS_{fh}' + S_{fh}B'F') + \lambda_{hh} F V_\beta F'](\mathrm{E} g_t g_t')^{-1}.$$

Proof. The proof is similar to that of the proof of Theorem 4.1 in West and McCracken (1998). □

V may be estimated using the usual techniques to account for serial correlation, including heteroskedasticity and autocorrelation consistent covariance matrices.

References

Chong, Y.Y., Hendry, D.F., 1986. Econometric evaluation of linear macro-economic models. Review of Economic Studies 53, 671–690.

Clemens, R.T., 1989. Combining forecasts: a review and annotated bibliography. International Journal of Forecasting 5, 559–583.

Cooper, R.L, 1972. The predictive performance of quarterly econometric models of the United States. In: Hickman, B (Ed.), Econometric Models of Cyclical Behavior. Vol. II. Columbia University Press, New York, pp. 813–925.

Cox, D.R., 1961. Tests of separate families of hypotheses. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Vol. I, University of California Press, Berkeley, CA, pp. 105–123.

Cox, D.R., 1962. Further results on tests of separate families of hypotheses. Journal of the Royal Statistical Society Series B 24, 406–424.

Davidson, R., MacKinnon, J.G., 1981. Several tests for model specification in the presence of alternative hypotheses. Econometrica 49, 781–795.

Diebold, F.X., 1989. Forecast combination and encompassing: reconciling two divergent literatures. International Journal of Forecasting 5, 589–592.

Diebold, F.X., 1998. Elements of Forecasting. Southwestern College Publishing, Cincinnati.

Engle, R.F., Che-Hsiung Hong, Kane, A., 1990. Valuation of Variance Forecasts with Simulated Options Markets, NBER Working Paper No. 3350.

Ericsson, N.R., 1992. Parameter constancy, mean square forecast errors, and measuring forecast performance: an exposition, extensions, and illustration. Journal of Policy Modeling 14, 465–495.

Ericcson, N.R., Marquez, J., 1993. Encompassing the forecasts of US trade balance models. Review of Economics and Statistics 75, 19–31.

Fair, R.C., Shiller, R., 1990. Comparing information in forecasts from econometric models. American Economic Review 80, 375–389.

Ferreira, S., 1999. GARCH interest rate volatility: a cross-country empirical test. Manuscript, University of Wisconsin.

Godfrey, L.G., 1998. Tests of non-nested regression models: some results on small sample behaviour and the bootstrap. Journal of Econometrics 84, 59–74.

Godfrey, L.G., Pesaran, M.H., 1983. Tests of non-nested regression models: small sample adjustments and Monte Carlo evidence. Journal of Econometrics 21, 133–154.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Kitamura, Y., 1997. Comparing misspecified dynamic econometric models using nonparametric likelihood. Manuscript, University of Wisconsin.

McAleer, M., 1995. The significance of testing empirical non-nested models. Journal of Econometrics 67, 149–171.

Mizon, G.E., Richard, J.F., 1986. The encompassing principle and its application to testing non-nested hypotheses. Econometrica 54, 657–678.

Stumborg, B., 1999. Estimating the impact of land use regulations on recreational property values in northern wisconsin. Manuscript, University of Wisconsin.

West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.

West, K.D., McCracken, M.W., 1998. Regression based tests of predictive ability. International Economic Review 39, 817–840.

White, H., 1994. Estimation, Inference and Specification Analysis. Cambridge University Press, Cambridge, UK.