MONEY DOES GRANGER-CAUSE OUTPUT IN THE BIVARIATE MONEY-OUTPUT RELATION*

Lawrence J. CHRISTIANO

Federal Reserve Bank of Minneapolis, Minneapolis, MN 55480, USA

Lars LJUNGQVIST

University of Minnesota, Minneapolis, MN 55455, USA
Federal Reserve Bank of Minneapolis, Minneapolis, MN 55480, USA

ABSTRACT

A bivariate Granger-causality test on money and output finds statistically significant causality when data are measured in log levels, but not when they are measured in first differences of the logs. Bootstrap simulation experiments indicate that, most probably, the first difference results reflect lack of power, whereas the level results reflect Granger-causality that is actually in the data. The reason for the lack of power in the first difference F-statistic is that first differencing the data appears to entail a specification error. By showing that money does Granger-cause output in the bivariate relation, we remove a potential embarrassment for models that assign an important role to money in business fluctuations.

Running Title: Granger-causality in the money-output relation

*We have benefited from the comments of Patrick Kehoe, Danny Quah, David Runkle, Thomas Sargent, Christopher Sims, James Stock, Richard Todd, Mark Watson and Ken West. We are particularly grateful for detailed comments from Martin Eichenbaum and to Stock for pointing out an error in the M1 data used in a previous version of this paper. Christiano gratefully acknowledges the financial support of the National Science Foundation. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

Send Galley Proofs to:

Lawrence J. Christiano
Research Department
Federal Reserve Bank of Minneapolis
Minneapolis, MN 55480

JEL Category #s

132, 211

1. An empirical puzzle

When we tested the null hypothesis that money fails to Granger-cause output in a bivariate money-output relation using data in log levels, the resulting F-statistic was 3.19 with significance level 0.0027. When instead we used first differences of the logged data, the resulting F-statistic was 1.38, with significance level 0.22.[1] Which of these two results is the most plausible--the first difference result, which suggests that money fails to Granger-cause output, or the level result, which suggests that money strongly Granger-causes output?

This empirical puzzle attracted our attention because of an argument in Eichenbaum and Singleton (1986). They conjecture that it would be difficult to construct a business cycle model which (a) assigns an important role to monetary factors, (b) is empirically plausible and (c) has the implication that money fails to Granger-cause output in the bivariate money-output relation.[2] If the Eichenbaum-Singleton conjecture is right, then a finding that there is no Granger-causality from money to output in the bivariate money-output relation would have major implications for monetary business cycle models.[3] For this reason, we thought it vital to resolve the empirical puzzle mentioned in the first paragraph.

Based on bootstrap simulation experiments, we find that the most likely explanation of the puzzle is that the small F-statistic based on the difference data reflects not the data's lack of Granger-causality from money to output, but rather the test's lack of power to detect it. The large F-statistic on the level data appears to reflect the greater power of this test to detect the Granger-causality that is in fact there. The reason for the lack of power of the first difference F-statistic is that, for our data set, log first differencing both time series prior to doing the Granger-causality test appears to give rise to specification error. Thus, we conclude that the bivariate Granger-causality pattern between money and output does not constitute an embarrassment to monetary business cycle models.

Our results would not be economically interesting if the Granger-causality from money to output was statistically significant, but numerically very small. Therefore, we also calculated the percent variance in the log of output that is due to a shock in money. Based on such a decomposition, we found that the Granger-causality from money to output is quantitatively substantial.

2. Possible explanations of the empirical puzzle

We use monthly observations on M1 and industrial production as our measures of money and output, respectively. The maintained hypothesis throughout this paper is that the joint log M1 (LM1) and log industrial production (LIP) process can be approximated by a seven-lag, bivariate vector autoregression (VAR) with a constant term. We examined the ability of four alternative sets of restrictions on this VAR to explain the empirical puzzle that motivates this paper: the high empirical level F-statistic and the low empirical difference F-statistic. We denote these sets of restrictions by H_O^D, H_A^D, H_O^L, H_A^L. Here,

H_O^D = Hypothesis that the level VAR can be represented as a six-lag bivariate VAR in first differences of LM1 and LIP, which satisfies the null hypothesis (H_O) that LM1 fails to Granger-cause LIP, i.e., fails to enter the LIP equation.

H_A^D = Same as H_O^D, except that the alternative hypothesis (H_A) is right and LM1 enters the LIP equation.

H_0^L and H_A^L = Analogous to H_0^D and H_A^D, except that the VAR in levels of LM1 and LIP is presumed not to be representable as a VAR in first differences of LM1 and LIP.

Two of these hypotheses, H_0^D and H_A^L, are of particular interest to us. Hypothesis H_0^D seemed plausible on a priori grounds for two reasons. First, if H_0^D is true, then the small empirical difference F-statistic is to be expected. Second, if H_0^D is true, then the level VAR representation of LM1 and LIP has two unit roots and LM1 and LIP are not cointegrated. Results in Sims, Stock and Watson (1986) indicate that, in this case, the sampling distribution of the level F-statistic is nonstandard. Although their results do not predict that application of standard asymptotic sampling theory necessarily leads to too many rejections (i.e., a high level F-statistic), this is a possibility.

Hypothesis H_A^L also appealed to us. First, in this case the high level F-statistic is to be expected. Second, under H_A^L the difference model is misspecified, and we thought it possible that this might lead to a small difference F-statistic. As we show below, our results indicate that H_A^L is the most likely explanation for the empirical puzzle.

3. Bootstrap simulation methodology

This section describes our methodology for investigating the empirical validity of the four hypotheses defined in section 2. We do so by examining the distribution of the level and difference F-statistics implied by each hypothesis. In addition, our analysis requires deriving the sampling distribution of the likelihood ratio statistic for testing the null hypothesis H_A^D against the alternative H_A^L. The distributions of these statistics were computed based on bootstrap simulations of four models, or data generating mechanisms (DGMs). We denote the DGM corresponding to H_i^j by DGM_i^j, for i = 0, A and j = D, L. We now describe the DGMs formally:

DGM_0^D = Bivariate, six-lag VAR in first differences of LM1 and LIP estimated using the 448 monthly observations from September 1948 to December 1985 as the estimation period and the March-August 1948 data as initial conditions. Each of the two equations in the VAR was estimated by ordinary least squares (OLS), and LM1 was restricted not to enter the LIP equation.

DGM_A^D = Same as DGM_0^D, except LM1 is permitted to enter the LIP equation with nonzero coefficients.

DGM_0^L = Bivariate, seven-lag VAR in levels of LM1 and LIP, estimated by OLS using the 448 monthly observations from September 1948 to December 1985 with data for February-August 1948 as initial conditions. The LIP equation was restricted not to include LM1.

DGM_A^L = Same as DGM_0^L, except LM1 was permitted to enter the LIP equation with nonzero coefficients.

Each DGM was used to generate 5,000 samples of 448 artificial observations on LM1 and LIP. Each DGM requires seven initial observations on LM1 and LIP to generate a sample. In all cases, we used the actual February-August 1948 data for this. In addition, two sets of 448 disturbances--one for each of the LIP and LM1 equations--are required. We call our simulations bootstrap simulations because these disturbances were obtained by randomly sampling from the fitted residuals computed during estimation of the given DGM.[4] For each DGM, we computed 5,000 difference and level Fs as follows.

Let H_0 denote the null hypothesis that LM1 fails to Granger-cause LIP, i.e., does not enter the LIP equation. On each of the 5,000 data sets, the level F and the difference F are the F-statistics for testing H_0 based on levels and differences of the data, respectively. These calculations were done in the same way as those underlying the empirical F-statistics reported in the introduction. In addition, the 5,000 data samples generated by DGM_A^D were used to compute 5,000 values of the likelihood ratio statistic for testing H_A^D against H_A^L.

As far as we know, the literature does not contain a formal justification for our bootstrap simulation methodology. We conjecture that there is an asymptotic justification. Our intuition is that with a large number of observations, consistency of the parameter estimates guarantees they are close to their true values, justifying centering the simulations on the estimated parameter values. For the same reason, we expect that in large samples the fitted disturbances resemble the true underlying disturbances, justifying sampling from the fitted disturbances.

We did two robustness checks on all our results in this paper. We redid all the calculations by drawing the disturbances from the normal distribution and found the results virtually unchanged. In addition, we redid our entire analysis using data for the period from January 1959 to December 1985 and obtained results very similar to those reported here. For details, see Christiano and Ljungqvist (1987).

4. The Granger-causality from money to output is statistically significant

We reached our conclusion that the Granger-causality from money to output is statistically significant based on the marginal distribution of the simulated difference and level Fs, which are graphed in figs. 1-4. The figures allow us to reject H_0^D and H_0^L at the 3% significance level.

Consider fig. 1, which is produced under H_0^D. In addition to reporting the frequency distribution of the simulated difference and level Fs, it also plots the F-distribution with 6 numerator and 435 denominator degrees of freedom. Conventional asymptotic sampling theory says this is the density function from which the empirical difference F-statistic is drawn under H_0^D, which is valid by construction in the simulations.

Note first in fig. 1 that the distribution of the simulated first difference F-statistics closely coincides with that of the theoretical Fs. This indicates that with 448 observations asymptotic theory is a pretty good approximation--an encouraging result in view of our belief that the justification for our methodology is asymptotic. Note also the fact that the simulated level Fs are only slightly shifted to the right of the simulated difference Fs. This means that, while a level F-statistic does tend to reject the null hypothesis too often if the difference model is right and money in fact does not Granger-cause output, the tendency is quantitatively too small to account for the high empirical F-value of 3.19 obtained using level data. In our 5,000 simulations, only 2.6% of the simulated level Fs exceed 3.19.

Fig. 2 shows that the empirical F-statistics--3.19 from levels and 1.38 from differences--are quite plausible under the H_A^L. The small magnitude of the empirical F-statistic based on differences is not surprising under the hypothesis that the data are generated by the level model. In our 5,000 simulations, for example, 16.5% of the difference F-statistics are even smaller than the empirical difference F of 1.38. The empirical level F of 3.19 is obviously also plausible relative to the simulated level F-statistics.

Fig. 3 allows us to assess the plausibility of H_0^L. This is a natural hypothesis to investigate given the results in fig. 2. One would like to know whether the large level F-statistics in that figure reflect the test's power or simply its tendency to reject too often, regardless of the status of the null hypothesis. Fig. 3 shows that the results in fig. 2 do reflect the power of the level F-statistic. In addition to reporting the frequency distribution of level and difference Fs under H_0^L, fig. 3 also reports the density function of the F-distribution with 7 numerator and 433 denominator degrees of freedom. Note that the three distributions in fig. 3 nearly coincide. We conclude that H_0^L can be rejected on the basis of the large empirical level F-statistic.

Finally, we turn to H_A^D. The frequency distribution of the level and difference Fs under this null hypothesis appear in fig. 4. The interesting thing to note is that both empirical F-statistics are individually plausible under H_A^D. For example, 18% of the difference Fs are smaller than the empirical difference F-value of 1.38. Also, 35.2% of the level Fs exceed the empirical F-value of 3.19.

Based on the analysis of the marginal distributions of the F-statistics, we conclude that H_0^D and H_0^L are not plausible, as asserted in the title of this section.

5. The empirical puzzle reflects that the difference model is misspecified

By focusing on the marginal distribution of the level and difference F-statistics, the preceding section was able to establish that the null hypothesis--LM1 fails to Granger-cause LIP--is implausible. However, our analysis did not allow us to conclude which model specification is the better one: the unrestricted first difference specification (H_A^D) or the unrestricted level specification (H_A^L). In the first part of this section we show that when the joint distribution of the Fs is considered, then of all the explanations cited in section 2, H_A^L is the most plausible. Based on a one-tailed and a two-tailed hypothesis test, we find that the others are rejected at the 5%

significance level. The second part of this section reports the results of a more conventional test of the first difference specification versus the level specification. There we show that the likelihood ratio test rejects the first difference specification at any positive significance level, thus corroborating the results in the first part of this section.

5.1. Tests based on the joint distribution of the level and difference F-statistics

To make our discussion of the hypothesis tests performed below precise, we first require some notation. Let $F_i^j(r)$ be a 2 × 1 vector with its first element containing a level F-statistic and its second element a difference F-statistic, generated at the rth simulation by data generating mechanism DGM_i^j, $i = 0, A$; $j = D, L$; $r = 1, \ldots, 5{,}000$. For example, $F_A^D(10)$ denotes the vector of F-statistics generated on the 10th simulation by the first difference VAR in which LM1 enters the LIP equation. Then write

$$F_i^j = \frac{1}{5{,}000} \sum_{r=1}^{5{,}000} F_i^j(r)$$

$$V_i^j = \frac{1}{5{,}000} \sum_{r=1}^{5{,}000} [F_i^j(r) - F_i^j][F_i^j(r) - F_i^j]'$$

where $i = 0, A$ and $j = D, L$. Thus, F_i^j is the 2 × 1 vector of means of the level and difference Fs from 5,000 simulations from model DGM_i^j, and V_i^j is the corresponding variance covariance matrix. These are reported in table 1. The second moment matrices in table 1 contain a correlation on the lower diagonal, a covariance on the upper diagonal and variances along the diagonal.

5.1.1. A one-tailed test

Note from table 1 that, in each case, the level and difference Fs are positively correlated. Interestingly, the highest positive correlation, 0.90, occurs when the data generating mechanism is DGM_A^D. Recall fig. 4's implication that under H_A^D the low empirical difference F and the high empirical level F are individually plausible and that the marginal distributions of the two Fs roughly coincide. Under these circumstances, the high positive correlation between the two suggests that the magnitude of the spread between the two empirical F-statistics is unlikely. This is confirmed by the results in table 2, which reports the frequency of the event $\{F_i^j(r)$: level $F > 3.19$, difference $F < 1.38\}$ for $j = D, L$ and $i = 0, A$. The table shows that among the $\{F_A^D(r), r = 1, \ldots, 5{,}000\}$ generated by DGM_A^D, only four were characterized by a difference F lower than the empirical one and a level F larger than the empirical one. Thus, relative to H_A^D, the empirical Fs lie far out in a tail, so that this test rejects H_A^D at the 0.1% significance level.

Table 2 also indicates that the empirical spread between the empirical F-statistics is very implausible relative to H_0^D and H_0^L. The hypothesis that comes out looking best by this test is H_A^L, which fails to be rejected at the 5% significance level. The plausibility of H_A^L is not overwhelming, however, since it is rejected at the 6% significance level by this test.

5.1.2. A two-tailed test

Further information about the distribution of the $F_i^j(r)$s is reported in figs. 5-8. These depict the scatter diagram of the level and difference Fs generated by each DGM. The empirical Fs are also reported in each graph. In addition, the two ellipses in the figures are confidence ellipsoids.[5] The smaller ellipse contains 90% of the realized $F_i^j(r)$s, and the larger one contains 95%. They are two-dimensional generalizations of the one-dimensional, two-sided confidence interval, symmetric about the mean. They provide an indication of the dispersion of the simulated F-statistics and form a basis for a second test of our hypotheses.

The confidence ellipsoids in the figures can be used to perform a two-tailed test. Doing so, we find that all our explanations for the empirical F-statistics are rejected at the 5% significance level, except H_A^L. Under H_0^D, H_A^D, H_0^L, H_A^L the empirical F-statistics lie on the boundary of the 96.1, 95.2, 98.6 and 39.7% confidence ellipsoids, respectively. (These ellipsoids are not depicted in the figures.) Clearly, this test is particularly favorable to H_A^L.

5.2. A likelihood ratio test

A more conventional way to compare H_A^D and H_A^L is to carry out a likelihood ratio test of the null hypothesis that the first difference VAR is true versus the alternative that the level VAR is true. Specifically, the empirical value of our test statistic is $\lambda = T \log \left[\det(V_D)/\det(V_L)\right]$, where $T = 448$ is the number of observations, V_D is the estimated innovation covariance matrix of the DGM_A^D model and V_L is the corresponding quantity for DGM_A^L. The value of λ is 39.24. If the model were covariance stationary under the null hypothesis, then--under the null hypothesis--λ would be a realization from a chi-square distribution with four degrees of freedom [see, e.g., Sims (1980a, p. 17)]. The number of degrees of freedom reflects the fact that there are four extra free parameters under the alternative hypothesis of the test. The significance level of 39.24 under a chi-square distribution with four degrees of freedom is approximately zero, so that application of conventional asymptotic sampling theory results in overwhelming rejection of the null hypothesis. However, this conclusion is premature, since the conventional asymptotic theory requires covariance stationarity under the null hypothesis--an assumption not satisfied in our case. As a result, we computed the significance level of our test statistic by bootstrap simulation. Specifically, we computed 5,000 artificial values of λ, one for each of the 5,000 data sets generated from DGM_A^D, as described in section 3. We found that not one of the 5,000 simulated λs exceeds 39.24, allowing us to reject H_A^D in favor of H_A^L at the zero significance level. Although we found that application of the conventional chi-square sampling theory would lead one to reject the null hypothesis too often if the data generating mechanism is DGM_A^D, the effect of this bias was obviously not quantitatively large enough to prevent overwhelming rejection of the null hypothesis.

We were initially surprised by this strong rejection, since informal evidence suggested to us that the two models are in fact similar. For example, the determinant of the autoregressive part of DGM_A^L has two roots, each of which appeared to us to be close to unity--1.003 and 0.991--raising the possibility that DGM_A^L is almost a first difference model.[6] In the appendix we show that the reason for the strong rejection of the difference model is the 1.003 root in the level model and the fact that, in the likelihood ratio sense, 1.003 is very far from unity and 0.991 is very close to it.

6. Power comparisons between the level and difference F-statistics

Informal comparison of figs. 2 and 3 suggests that the level F has greater power when the level specification is correct. At first glance, fig. 4 might be interpreted as saying that the level F-statistic also has greater power when the first difference specification is correct. However, recall that power is defined as the probability of rejecting a false null hypothesis given a fixed probability of rejecting the null hypothesis when it is true (i.e., committing a Type I error). Fig. 1 shows that the level F has a tendency to reject the null hypothesis more often than the first difference F when the null hypothesis is true and the data are generated by the difference model. The fact that the horizontal distance between the level and first difference F is greater in fig. 1 than in fig. 4 suggests that the power of the first difference F-statistic may exceed that of the level F when the first difference model is true. These observations are confirmed by the results in table 3.

Table 3 shows that the power of the level F-statistic substantially exceeds that of the difference F-statistic when the null hypothesis (H_0) is characterized by DGM_0^L and the alternative hypothesis is characterized by DGM_A^L. For example, given a Type I error probability of 5%, the level F correctly rejects H_0 93.66% of the time, whereas the difference F rejects only 51.98% of the time. When the difference specification is correct, then the difference F-statistic is the more powerful one. This is perhaps not surprising since the VAR parameter estimators that underlie the difference F-statistic are more efficient, in this case, than the estimators underlying the level F-statistic.

7. The Granger-causality from money to output is quantitatively substantial

Finally, we investigated whether the Granger-causality from LM1 to LIP is strong enough to deserve attention. We calculated the percent of the conditional variance in LIP that is due to a shock in LM1 at various horizons. These calculations require normalizing the disturbance vector. We chose to do so by restricting money disturbances not to affect LIP in the current month. We expect that this choice of normalization does not affect the results because the correlation between the innovations to LM1 and LIP is close to zero.[7] We did the calculations based on the unrestricted level model (DGM_A^L) and the unrestricted difference model (DGM_A^D).

Our results based on using the level VAR are reported in table 4. It indicates that, in the bivariate relation, innovations to the log of money play a numerically important role in explaining variations in the log of output. In a 70% confidence interval, their importance exceeds 14% at the two-, three-, and four-year horizons, and the corresponding mean estimate is between 25 and 30%.

We also investigated the implications of the unrestricted first difference VAR model for the decomposition of variance in the log first difference of output. According to that model, at the 12-month horizon, 4.88% (2.17-7.64%) of the variance in output growth is due to innovations in money; at the 24-month horizon, 5.15% (2.27-8.06%). (Numbers in parentheses are the 70% confidence interval.) The variance decomposition (and 70% confidence interval) converges by the 24th month. These calculations were done in the same way--with obvious modifications--as those for table 4.[8] However, the results are not directly comparable since they pertain to the variance decomposition of the first difference of the log of output, whereas those in table 4 pertain to the log of output.

Table 5 reports the variance decomposition for the log level of output implied by the estimated unrestricted first difference VAR.[9] The mean estimates in that table also support our contention that the Granger-causality from money to output is quantitatively large. However, the confidence intervals are shifted closer to zero relative to those in table 4. This presumably, is another manifestation of the first difference model's tendency to regard the Granger-causality from money to output as statistically insignificant. Recall from section 5 that this implication of the difference model is implausible given the magnitude of the spread between the empirical difference and level F-statistics. In addition, we showed in that section that the first difference representation is strongly rejected in favor of the level representation.

8. Summary and conclusion

When a regression test of the null hypothesis (H_0) that money fails to Granger-cause output is carried out using first differences of the data, the resulting test statistic is small, apparently indicating no evidence against H_0. When the test is carried out using levels of the data, the test statistic is large, apparently indicating rejection of H_0. Bootstrap simulations showed that a VAR fit to first differences of the data could not account for the simultaneously large levels test statistic and small first difference test statistic. This was true whether or not the estimated VAR in differences was restricted to exclude money from the output equation. We also showed that a VAR based on levels of the data--restricted so that money does not enter the output equation--could not account for the empirical test statistics. One model which could account for these results is the VAR estimated using levels of the data without imposing any restrictions. We showed that this latter model implies that the Granger-causality from money to output is quantitatively large. Based on these results and the fact that the first difference model is strongly rejected by a likelihood ratio test, we conclude that Granger-causality from money to output is statistically and quantitatively significant. First differencing prior to executing the Granger-causality test appears to have resulted in a substantial loss of power of that test. This reduction in power due to first differencing reflects that, for this data set, first differencing both series seems to involve a specification error.

As with any empirical study, one has to bear in mind that our results are based on a particular maintained hypothesis. The maintained hypothesis is that the (log) levels of output and money have a seven-lag VAR representation. The work of Eichenbaum and Singleton (1986) and Stock and Watson

(1987) suggests that a possible source of misspecification of our model lies with our exclusion of a time trend. This possibility is worth further investigation. However, in the meantime, two points are worth stressing. First, Stock and Watson (1987) show that our basic result--that the Granger-causality from money to output is significant--is in any case robust to possible misspecification along this dimension. Second, the work of Quenouille (1947) and Chow and Levitan (1969) suggests that, due to the explosive root in our DGM_A^L model, it looks in some respects like a model with a time trend anyway.[10]

We hope not only that this paper sheds light on the nature of money-output dynamics, but that it is of more general methodological interest. First, we think it illustrates the power and versatility of bootstrap simulations for conducting inference in contexts where the required sampling theory either is intractable or requires extensive specialized knowledge of econometrics. In particular, we have applied it in cases where there are unit roots and explosive roots and where the underlying model is misspecified. The latter plays a central role in the context of non-nested tests and encompassing tests [see, e.g., Mizon and Richard (1986) and the references they cite]. Second, the analysis of data with VARs has at times been criticized for its apparent lack of robustness to whimsical assumptions, such as whether or not the data have been first differenced. By presenting one example in which such an apparent lack of robustness is decisively resolved, this paper makes us hopeful that our methodology can do the same in other such cases.

Notes

[1] The level F-statistic is based on a regression of the log of industrial production (LIP) on a constant, seven lags of that log and seven lags of the log of M1 (LM1). Similarly, the difference F-statistic is based on a regression of the first difference of LIP on a constant and six lags of first differences of LIP and LM1. The estimation period is from September 1948 to December 1985, and the initial conditions are the February-August 1948 observations. The significance level of the test is the area under the F-distribution to the right of the computed test statistic.

[2] It is straightforward to produce monetary models which satisfy only (a) and (c) and not (b). Mankiw (1986) has two such examples. His first example implies that the regression of output on money has an R^2 of unity. His second example assumes that the money stock is a white-noise process. Both of these specifications appear to be empirically unreasonable. [For an extended discussion and further examples, see Eichenbaum and Singleton (1986, sect. 4b).]

[3] It is important to emphasize that our results pertain to the <u>bivariate</u> relationship between money and output. For example, it is well known that when a financial rate of return is included in the VAR, then money fails to Granger-cause output, even in a level specification [Sims (1980b)]. Sims (1980b, sect. III) shows that there are ways to reconcile this finding with the view that monetary policy plays an important role in business fluctuations. However, these potential explanations presume that money Granger-causes output in the bivariate system. It is this assumption that is at issue in this paper.

[4] This was done as follows. Let $\hat{\varepsilon}_t$, $t = 1, \ldots, 448$, be the set of 2×1 vectors of fitted disturbances from the estimated VAR model. One draw,

$\{\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \ldots, \tilde{\varepsilon}_{448}\}$, was executed by first randomly selecting 448 integers, i_1, i_2, ..., i_{448}, with replacement, from $\{1,2,\ldots,448\}$. Then $\tilde{\varepsilon}_t \equiv \hat{\varepsilon}_{i_t}$, $t = 1$, ..., 448.

[5] The confidence ellipsoids were computed as follows. Let $s_i^j[x]$ be a function mapping R_+^2 into R_+^1 defined as follows:

$$s_i^j[x] = (x-F_i^j)(V_i^j)^{-1}(x-F_i^j)'$$

for $i = 0, A$; $j = D, L$. Also, let $c_i^j(\alpha)$ be an $\alpha\%$ critical value, defined by the property $s_i^j[F_i^j(r)] \geq c_i^j(\alpha)$ for no more than $(\alpha/100) \times 5{,}000$ values of $r \in \{1, \ldots, 5{,}000\}$. Then the $\alpha\%$ confidence ellipsoid is defined as the set of points x such that

$$s_i^j(x) = c_i^j(\alpha)$$

for $j = D, L$; $i = 0, A$.

[6] The roots of the determinant of the autoregressive part of the DGM_A^L model are $-0.475 \pm 0.593i$, -0.504, -0.475, $-0.128 \pm 0.575i$, $0.411 \pm 0.586i$, 0.991, 1.003, $0.695 \pm 0.156i$ and $0.284 \pm 0.187i$. The corresponding roots of the DGM_A^D model are $-0.472 \pm 0.621i$, -0.438, -0.576, $-0.127 \pm 0.555i$, $0.394 \pm 0.639i$, 0.864, 1.0, 1.0, $0.558 \pm 0.153i$ and 0.119. These roots are defined as follows. Let $A(L)y_t = \varepsilon_t$ be a bivariate, p^{th} order VAR, where ε_t is white noise and uncorrelated with y_{t-s}, $s > 0$; $A(L) = I - A_1L - A_2L^2 - \ldots - A_pL^p$ and L is the lag operator. Denote the ij^{th} polynomial element of A(L) by $a_{ij}(L)$ for $i, j = 1, 2$. Then the roots are the reciprocals of z values such that $\det[A(z)] = a_{11}(z)a_{22}(z) - a_{12}(z)a_{21}(z) = 0$.

[7] The correlation between the residuals in the LIP and LM1 equation is 0.065 for the model specified in levels (DGM_A^L) and 0.025 for the model specified in first differences (DGM_A^D).

[8] The difference in the way the calculations were done is the following: The data generating mechanism underlying the results in table 4 is the estimated unrestricted VAR in levels (DGM^L_A), whereas the data generating mechanism underlying the calculations just described is the estimated VAR in first differences (DGM^D_A).

[9] The calculations underlying table 5 were done in a similar manner as those underlying table 4. In particular, 5,000 artificial data sets were generated by the DGM^D_A model in the manner described in section 3. On each data set an unrestricted, six-lag, bivariate VAR in first differences was estimated. Variance decompositions of the log of output were then computed based on the parameter values of the implied levels VAR model. An alternative strategy for computing the distribution of the variance decompositions would have exploited the asymptotic normality of the parameter estimators of the first difference VAR, as in Christiano (1986). We chose to use our more computer-intensive approach to preserve symmetry with the calculations underlying table 4. In the latter case, we were not happy with using standard asymptotic normality results because the level VAR specification appears not to be covariance stationary.

[10] Stock and Watson (1987) note that the first difference of LIP appears not to exhibit a trend, whereas the first difference of LM1 does. They measure the importance of the trend in a variable by the magnitude of the t-statistic on the coefficient on time in the regression of the growth of the variable on six lags of its growth rate, a constant and time. These t-statistics for LIP and LM1 based on our data sets are -0.496 and 4.89, respectively. We computed these same t-statistics in each of the 5,000 artificial data sets generated by DGM^L_A and found the average t for LIP and LM1 to be -0.413 (55.28%) and 4.97 (54.74%), respectively. (Numbers in parentheses are

the percent of simulated t-statistics that exceeded the corresponding empirical value.) Thus, consistent with the analyses of Quenouille (1947) and Chow and Levitan (1969) and the simple example in Sargent (1979, p. 293), our model seems capable of capturing the trend behavior of the LM1 and LIP data as measured by Stock and Watson (1987).

Appendix:

Sensitivity of the likelihood ratio statistic to an explosive root

In section 5.2 we showed that the null hypothesis that DGM_A^D is the data generating mechanism is rejected at any positive significance level against DGM_A^L. We also noted that the two maximal roots of the determinant of the autoregressive part of DGM_A^L are 0.991 and 1.003 (see footnote 6). The rejection of the DGM_A^D model is due to the discrepancy between these roots and unity in the sense that if they had been unity instead, then the DGM_A^L would be a first difference model. This appendix also shows that, of these two roots, it is the explosive one (1.003) which principally accounts for the overwhelming nature of the rejection reported in section 5.2. In particular, if the 1.003 root had been 1.000 instead, then most likely we would not have rejected the first difference model. This implies that the next smaller root in DGM_A^L, 0.991, plays very little, if any, role in accounting for the rejection of the first difference model. Thus an explosive root and its reciprocal have very different impacts on inference. We give precise measures of this asymmetry using power calculations.

Write the DGM_A^L as follows:

$$A(L)y_t = \varepsilon_t \tag{A.1}$$

where the constant is suppressed for notational simplicity and $y_t \equiv (LIP_t, LM1_t)'$ and $\varepsilon_1, \ldots, \varepsilon_{448}$ are the fitted disturbances. Also, $A(L) = I - A_1L - A_2L^2 - \ldots - A_7L^7$ are the estimated VAR parameters and L is the backshift operator; i.e., $L^i y_t \equiv y_{t-i}$ for any integer i. As we noted in the previous paragraph, the maximal root of $\det[A(z^{-1})]$ is 1.003 and the next smaller one is 0.991.

To establish that the rejection of the first difference specification results from the explosive root in A(L), we show that $\lambda = 39.24$ is very plausible under DGM_A^L, but very implausible relative to a perturbation of DGM_A^L in which the 1.003 root of A(L) is replaced by 1.000. The other maximal root, 0.991, plays at best a small role accounting for the high likelihood ratio statistic, since replacing it by 1.000 in A(L) has little effect on the plausibility of $\lambda = 39.24$. To make these observations precise, we need to explain what we mean by "replacing" roots in A(L) and what we mean by the "plausibility" of $\lambda = 39.24$ relative to a given model.

To explain how we "replaced" one or both of the maximal roots of A(L), we first require some notation. Let $\bar{\mu}_i$, i = 1, ..., 14, denote the zeroes of $\det[A(z^{-1})]$ with the convention $\bar{\mu}_1 = 0.991$ and $\bar{\mu}_2 = 1.003$. Similarly, let x^i denote the eigenvector of $A(\bar{\mu}_i{}^{-1})$, i = 1, ..., 14, and let X be the 2 × 2 matrix $[x^1x^2]$. In our case, X is nonsingular, so that X^{-1} exists. The polynomial matrix A(L) can be completely characterized in terms of the x^i's and $\bar{\mu}_i$'s. We obtained perturbations on A(L) by altering the values of its maximal roots without touching x^i, i = 1, ..., 14, or $\bar{\mu}_i$, i = 3, ..., 14. For $(\mu_1,\mu_2) \in R^2$, let $A(L;\mu_1,\mu_2)$ denote such a perturbation. It is uniquely represented as follows:

$$\text{(A.2)} \qquad A(L;\mu_1,\mu_2) = \tilde{A}(L)XG(L;\mu_1,\mu_2)X^{-1}$$

where $\tilde{A}(L) \equiv A(L)XG(L;\bar{\mu}_1,\bar{\mu}_2)^{-1}X^{-1}$ and

$$\text{(A.3)} \qquad G(L;\mu_1,\mu_2) = \begin{bmatrix} 1-\mu_1 L & 0 \\ 0 & 1-\mu_2 L \end{bmatrix}.$$

Also, $\tilde{A}(L)$ is a sixth order matrix polynomial in L with $\tilde{A}(0) = I$. The latter follows trivially from the fact that $A(0) = G(0;\bar{\mu}_1,\bar{\mu}_2) = I$. To see the former, note that $\tilde{A}(L)X = A(L)XG(L;\bar{\mu}_1,\bar{\mu}_2)^{-1}$ is a sixth order matrix polynomial in L since the i^{th} column of $A(L)X$ is proportional to $(1-\bar{\mu}_iL)$, i = 1, 2. This follows from the fact that the i^{th} column of $A(\bar{\mu}_i^{-1})X$ contains zeroes, i = 1, 2. It can be verified that $A(L;1.00,1.00) = (1-L)\tilde{A}(L)$, which is a first difference model for y_t.

We measure the plausibility of $\lambda = 39.24$ relative to a given VAR model by its proximity to the central tendency of the likelihood ratio test statistic, as implied by the model. To establish our claim that without a maximal root of 1.003 in A(L), $\lambda = 39.24$ is implausible, we require the density function for the likelihood ratio statistic implied by each of the three models: $A(L;1.000,1.003)$, $A(L;0.991,1.000)$ and $A(L) \equiv A(L;0.991,1.003)$. These were calculated based on 5,000 artificial data sets of 448 observations on y_t generated from each of the three models. The ingredients that went into each simulation were one of the above-mentioned sets of VAR coefficients, the actual February-August 1948 observations on y_t and a set of 448 disturbances, randomly sampled from ε_t, t = 1, ..., 448. On each artificial data set, we computed a likelihood ratio statistic using the same procedure as the one used to arrive at $\lambda = 39.24$ (see section 5.2). This gave us three frequency distributions of likelihood ratio statistics, which approximate the underlying density functions of interest and are plotted in fig. A1. The empirical value of the likelihood ratio statistic, 39.24, is indicated on the horizontal axis. The other two curves in fig. A1 are not relevant at this point, but play a role in our discussion below.

Note first from fig. A1 that $\lambda = 39.24$ is very plausible under DGM_A^L. [See the curve labeled A(L;0.991,1.003).] In particular, 71.46% of the simulated likelihood ratio statistics from this model exceed 39.24. When the 0.991 root in A(L) is replaced by unity, the frequency distribution of likelihood ratio statistics shifts only a little to the left [see the curve labeled A(L;1.00,1.003)], with the consequence that λ = 39.24 remains plausible. In this case, 30.64% of the simulated likelihood ratio statistics exceed 39.24. Thus, whether the second largest root in A(L) is 0.991 or 1.000 makes relatively little difference to the magnitude of the likelihood ratio statistic. This strongly contrasts with the quantitatively large role played by the explosive root. When the 1.003 root is replaced by 1.000, the distribution of likelihood ratio statistics shifts so sharply left that λ = 39.24 is extremely improbable. In particular, of the 5,000 artificial likelihood ratio statistics simulated from A(L;0.991,1.000), none exceed the empirical value of 39.24. Thus, as between the two maximal roots of A(L), the explosive one plays an essential role in accounting for the overwhelming rejection of the first difference model reported in 5.2.

Another way to measure the role of the two maximal roots of A(L) in accounting for λ = 39.24 examines their impact on the power of the likelihood ratio statistic to reject the first difference specification. Our power calculations appear in table A1. The calculations were based on the three sets of 5,000 artificial likelihood ratio statistics generated in the manner described above by A(L;1.000,1.003), A(L;0.991,1.003) and A(L;0.991,1.000). In addition, we obtained 5% and 10% critical values by simulating 5,000 likelihood ratio statistics using the A(L;1.00,1.00) model, in which the null hypothesis is true by construction. These simulations used random samples of ε_t, t = 1, ..., 448, and the February-August initial conditions on y_t. The frequency distribution of the simulated likelihood ratio statistics appears in fig. A1. It is worth noting that we also calculated this frequency distribution based on the 5,000 data samples generated by DGM_A^D as discussed in section 3 and found it to be virtually identical to the one implied by A(L;1.00,1.00). In addition, it is interesting to note that the frequency distribution implied by A(L;1.00,1.00) lies to the right of the chi-square distribution with four degrees of freedom. The latter is what blind application of covariance stationary, asymptotic sampling theory would have led to. In particular, the use of standard sampling theory leads to too frequent rejection of the null hypothesis.

The first row in table A1 shows that the likelihood ratio statistic has enormous power against A(L;0.991,1.003), the power being close to its upper bound of 100%. The second row in the table shows that when the 1.003 root of A(L) is replaced by 1.000, the power of the test drops precipitously--all the way to 23% with a 5% size. This loss of power reflects that μ_1 = 0.991 is hard to distinguish from μ_1 = 1.000, the value of μ_1 under the null hypothesis of the test. The third row in the table shows that when the failure of the null hypothesis is due only to μ_2 = 1.003, then the power is nearly as high as it is in the first row and is almost 100%. This establishes that the high power in the first row is due principally to the explosive root and has little to do with the 0.991 root. An informal way to summarize these results is that although the Euclidean distance implies that 1.003 is closer to unity than is 0.991, in a likelihood ratio sense 1.003 is very far from unity and 0.991 is very close to it.

References

Chow, G. C. and R. E. Levitan, 1969, Spectral properties of non-stationary systems of linear stochastic difference equations, Journal of the American Statistical Association 64, 581-590.

Christiano, L. J., 1986, Money and the U.S. economy in the 1980s: A break from the past?, Federal Reserve Bank of Minneapolis Quarterly Review 10, 2-13.

Christiano, L. J. and L. Ljungqvist, 1987, Technical appendix to "Money does Granger-cause output in the bivariate money-output relation," Research Department working paper 369 (Federal Reserve Bank of Minneapolis, Minneapolis, MN).

Eichenbaum, M. and K. J. Singleton, 1986, Do equilibrium real business cycle theories explain postwar U.S. business cycles?, in: S. Fischer, ed., NBER Macroeconomics Annual 1986 (MIT Press, Cambridge, MA) 91-135.

Litterman, R. B., 1979, Techniques of forecasting using vector autoregressions, Research Department working paper 115 (Federal Reserve Bank of Minneapolis, Minneapolis, MN).

Mankiw, N. G., 1986, Comment on Eichenbaum and Singleton, in: S. Fischer, ed., NBER Macroeconomics Annual 1986 (MIT Press, Cambridge, MA) 139-145.

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

Quenouille, M. H., 1947, The analysis of multiple time-series (Charles Griffin, London).

Sargent, T. J., 1979, Macroeconomic theory (Academic Press, New York).

Sims, C. A., 1980a, Macroeconomics and reality, Econometrica 48, 1-48.

Sims, C. A., 1980b, Comparison of interwar and postwar business cycles: Monetarism reconsidered, American Economic Review (Papers and Proceedings) 70, 250-257.

Sims, C. A., J. H. Stock and M. W. Watson, 1986, Inference in linear time series models with some unit roots, Manuscript (Research Department, Federal Reserve Bank of Minneapolis, Minneapolis, MN).

Stock, J. H. and M. W. Watson, 1987, Interpreting the evidence on money-income causality, Working paper 2228 (National Bureau of Economic Research, Cambridge, MA).

Table 1

First and second moments of simulated F-statistics.

Data generating mechanism[a]	Mean[b]	Correlation/variance matrix[c]	
DGM_O^D	1.51	0.51	0.31
	1.00	0.72	0.36
DGM_A^L	4.64	2.60	1.57
	2.53	0.81	1.43
DGM_O^L	1.18	0.41	0.25
	1.10	0.60	0.41
DGM_A^D	2.89	1.33	1.24
	2.50	0.90	1.43

[a] See section 3 for definitions of data generating mechanisms (DGMs).

[b] This is F_i^j, i = O, A and j = D, L. The first element of the vector pertains to the level F-statistic. See section 5.1 for details.

[c] This is V_i^j, i = O, A and j = D, L. This is the variance covariance matrix of the simulated level and difference Fs, except that the 2,1 element of the reported matrix is the correlation. The 1,1 element is the variance of the simulated level Fs. See section 5.1 for details.

Table 2

Frequency of event

(level F > 3.19, difference F < 1.38).

Model	Number of occurrences (out of 5,000)	Frequency
DGM_O^D	23	0.46%
DGM_A^L	280	5.60%
DGM_O^L	8	0.16%
DGM_A^D	4	0.08%

Table 3

Power comparisons of level and difference F.[a]

	Prob (Type I error) = 5%		Prob (Type I error) = 10%	
	Critical value	Power	Critical value	Power
Difference model[b]				
Difference F	2.11 (5.13%)[c]	58.18	1.78 (10.14%)	69.60
Level F	2.81 (0.71%)	48.18	2.43 (1.89%)	62.48
Level model[d]				
Difference F	2.31 (3.31%)	51.98	1.95 (7.10%)	64.14
Level F	2.42 (1.96%)	93.66	2.04 (4.93%)	96.98

[a] The table reports critical values and powers for the level and first difference F-statistics. The "critical value" columns report critical values for the indicated F-statistics that result in the null hypothesis (H_0) being rejected the indicated percent (5 or 10) of times when H_0 is true. Here, H_0 = LM1 does not Granger-cause LIP. The "power" columns report the percent of times H_0 is rejected using the indicated critical value for the F-statistic when the alternative hypothesis (H_A) is true.

[b] Under H_0, the data generating mechanism is DGM_0^D; under H_A, it is DGM_A^D.

[c] Numbers in parentheses are the area under the F(n,d) distribution to the right of the associated critical value. In the case of the difference F-statistic, n = 6, d = 435; in the case of the level F-statistic, n = 7, d = 433.

[d] Under H_0, the data generating mechanism is DGM_0^L; under H_A, it is DGM_A^L.

Table 4

Percent variance in the log of output due to
an orthogonalized disturbance in the log of money
in the unrestricted level model.[a]

			Confidence intervals	
Horizon (months)	Mean	Standard deviation	70%	90%
12	18.09	7.37	(10.40-25.91)	(6.97-31.41)
24	25.00	9.76	(14.66-35.39)	(10.01-42.07)
36	27.80	10.57	(16.63-39.14)	(11.37-46.08)
48	29.71	11.07	(17.99-41.55)	(12.25-48.62)

[a] These are the results of 5,000 simulated data sets generated by the estimated unrestricted level model in the manner described in section 3. On each data set, a seven-lag, bivariate VAR in levels was estimated. The parameter estimates were then used--via the formulas in Litterman (1979, p. 76)--to compute the percent variance in the log of output due to an orthogonalized innovation in the log of money for each of the indicated horizons. The mean and standard deviation values are the average and standard deviations of those quantities. An x% confidence interval is a pair of numbers--say, x_ℓ and x_u--such that x_ℓ is greater than and x_u is less than (100-x)/2% of the quantities.

Table 5

Percent variance in the log of output due to
an orthogonalized disturbance in the log of money
in the unrestricted first difference model.[a]

			Confidence intervals	
Horizon (months)	Mean	Standard deviation	70%	90%
12	6.99	4.70	(2.29-11.75)	(0.90-15.90)
24	13.14	8.66	(4.23-22.28)	(1.49-29.77)
36	16.05	10.49	(5.18-27.38)	(1.76-36.08)
48	17.58	11.46	(5.65-29.90)	(1.91-39.32)

[a] For an explanation, see footnote 9.

Table A1

Frequency of rejecting a false null hypothesis ($\mu_1 = \mu_2 = 1$).

Roots			
μ_1	μ_2	5% Size[a]	10% Size
0.991	1.003	99.94%	99.96%
0.991	1.000	22.58%	34.22%
1.000	1.003	98.10%	99.06%

[a] Frequency of times that the null hypothesis ($\mu_1,\mu_2 = 1$) is rejected using a critical value for the likelihood ratio statistic (λ defined in section 5.2) that results in rejecting the null hypothesis 5% of the time when it is true.

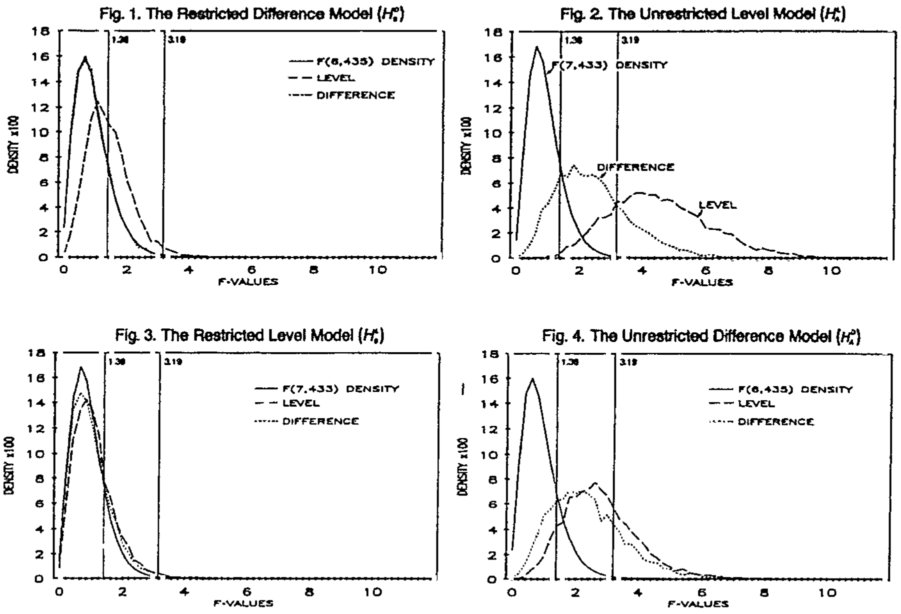


Fig. 1. The Restricted Difference Model (H_0^D)

Fig. 2. The Unrestricted Level Model (H_A^L)

Fig. 3. The Restricted Level Model (H_0^L)

Fig. 4. The Unrestricted Difference Model (H_A^D)

NOTE: Figure headings refer to the data-generating mechanism (*DGM*) used to generate 5,000 artificial data sets for the log of M1 (*LM*1) and the log of industrial production (*LIP*). Then DGM_0^D and DGM_A^D are estimated six-lag, bivariate VARs specified in the first difference of *LM*1 and *LIP*. DGM_0^L and DGM_A^L are estimated seven-lag, bivariate VARs specified in levels of *LM*1 and *LIP*. In DGM_0^D and DGM_0^L, the VARs are restricted so that *LM*1 does not enter the *LIP* equation, whereas DGM_A^D and DGM_A^L are estimated without restrictions. In each case, model estimates are consistent under the corresponding hypothesis H_k^j where $j = L, D$ and $k = 0, A$. Each figure depicts the frequency distribution of three *F*-statistics:

- Curves labeled *difference* are the frequency distribution of 5,000 *F*-statistics (one for each of the 5,000 data sets) for testing the null hypothesis that *LM*1 fails to Granger-cause *LIP* (H_0) based on a bivariate, six-lag VAR in the first difference of *LIP* and *LM*1.
- Curves labeled *level* are the frequency distribution of 5,000 *F*-statistics for testing H_0 based on a bivariate, seven-lag VAR in levels of *LIP* and *LM*1.
- *F(n,d) density* is the theoretical *F*-distribution with *n* numerator and *d* denominator degrees of freedom.

The numbers 1.38 and 3.19 are the empirical difference and level *F*-statistics, respectively.

Figs. 1-4. Frequency distributions of the theoretical and empirical level and difference F-statistics implied by four hypotheses: H_0^D, H_A^L, H_0^L, H_A^D.

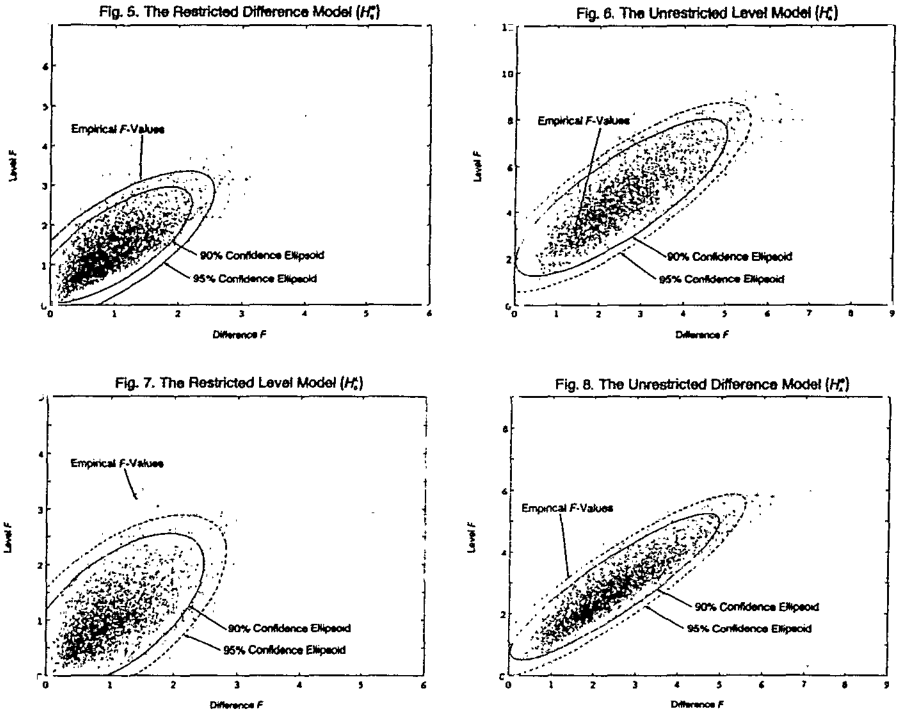


Fig. 5. The Restricted Difference Model (H_0^D)

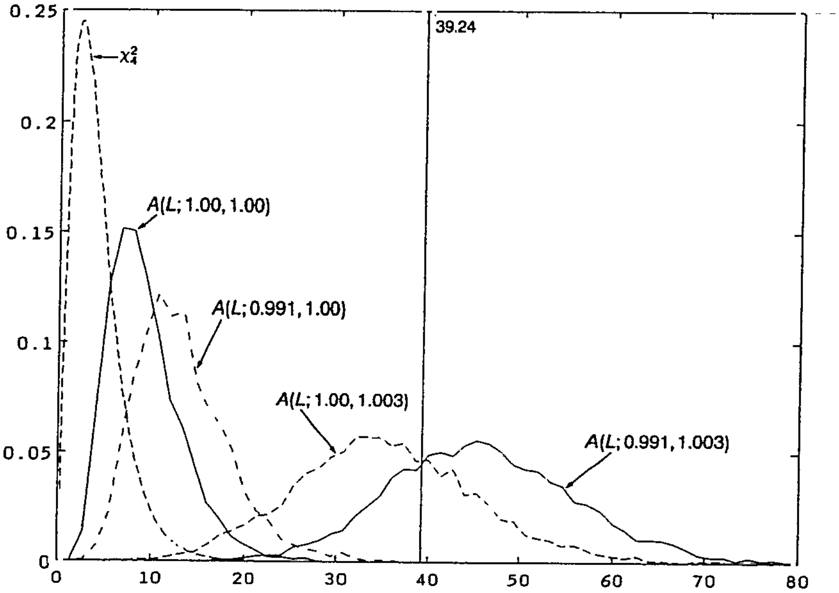
Fig. 6. The Unrestricted Level Model (H_A^L)

Fig. 7. The Restricted Level Model (H_0^L)

Fig. 8. The Unrestricted Difference Model (H_A^D)

NOTE: Figs. 4 + *i* contain scatterplots of the two simulated *F*-statistics whose marginal densities are reported in figs. *i*, for *i* = 1, 2, 3, 4. For further information on these, see the note to figs. 1-4. The *x*% confidence ellipsoids contain *x*% of the points in the scatterplots, for *x* = 90, 95. See section 5 and footnote 5 for a discussion of these. Finally, each figure also indicates the location of the empirical *F*-values. We conclude that the model underlying fig. 6 is the most plausible, since it is the only one which contains the empirical *F*s well within the interior of its scatterplot. (Note that the scales on these figures are different.)

Figs. 5-8. Scatterplots of the theoretical and empirical level and difference F-statistics implied by four hypotheses: H_0^D, H_A^L, H_0^L, H_A^D.



NOTE: The four curves labeled $A(L;\mu_1,\mu_2)$ for various values of μ_1,μ_2 are frequency distributions of the likelihood ratio statistic testing the null hypothesis of a six-lag bivariate VAR specification in differences versus the alternative of a seven-lag bivariate specification in levels. The fifth curve is the density function of the chi-square distribution with four degrees of freedom. The number 39.24 is the empirical value of the likelihood ratio statistic

Fig. A1. A chi-square distribution and four frequency distributions of likelihood ratio statistics testing the level and difference specifications.