Chapter 16

Multivariate: Vector Autoregression

The regression model is an explicitly multivariate model, in which variables are explained and forecast on the basis of their own history and the histories of other, related, variables. Exploiting such cross-variable linkages may lead to good and intuitive forecasting models, and to better forecasts than those obtained from univariate models.

Regression models are often called causal, or explanatory, models. For example, in the linear regression model,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

the presumption is that x helps determine, or cause, y, not the other way around. For this reason the left-hand-side variable is sometimes called the "endogenous" variable, and the right-hand side variables are called "exogenous" or "explanatory" variables.

But ultimately regression models, like all statistical models, are models of correlation, not causation. Except in special cases, all variables are endogenous, and it's best to admit as much from the outset. In this chapter we'll explicitly do so; we'll work with systems of regression equations called vector autoregressions (VARs).

16.1 Distributed Lag Models

An unconditional forecasting model like

$$y_t = \beta_0 + \delta x_{t-1} + \varepsilon_t$$

can be immediately generalized to the distributed lag model,

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t.$$

We say that y depends on a distributed lag of past x's. The coefficients on the lagged x's are called lag weights, and their pattern is called the lag distribution.

One way to estimate a distributed lag model is simply to include all N_x lags of x in the regression, which can be estimated by least squares in the usual way. In many situations, however, N_x might be quite a large number, in which case we'd have to use many degrees of freedom to estimate the model, violating the parsimony principle. Often we can recover many of those degrees of freedom without seriously worsening the model's fit by constraining the lag weights to lie on a low-order polynomial. Such polynomial distributed lags promote smoothness in the lag distribution and may lead to sophisticatedly simple models with improved forecasting performance.

Polynomial distributed lag models are estimated by minimizing the sum of squared residuals in the usual way, subject to the constraint that the lag weights follow a low-order polynomial whose degree must be specified. Suppose, for example, that we constrain the lag weights to follow a second-degree polynomial. Then we find the parameter estimates by solving the problem

$$\min_{\beta_0,\,\delta_i} \sum_{t=N_x+1}^{T} \left[y_t - \beta_0 - \sum_{i=1}^{N_x} \delta_i x_{t-i} \right]^2,$$

subject to

$$\delta_i = P(i) = a + bi + ci^2,\ i = 1, ..., N_x.$$

This converts the estimation problem from one of estimating $1 + N_x$ parameters, $\beta_0, \delta_1, ..., \delta_{N_x}$, to one of estimating four parameters, β_0, a, b and c. Sometimes additional constraints are imposed on the shape of the polynomial, such as $P(N_x) = 0$, which enforces the idea that the dynamics have been exhausted by lag N_x.

Polynomial distributed lags produce aesthetically appealing, but basically ad hoc, lag distributions. After all, why should the lag weights necessarily follow a low-order polynomial? An alternative and often preferable approach makes use of the rational distributed lags that we introduced in Chapter 7 in the context of univariate $ARMA$ modeling. Rational distributed lags promote parsimony, and hence smoothness in the lag distribution, but they do so in a way that's potentially much less restrictive than requiring the lag weights to follow a low-order polynomial. We might, for example, use a model like

$$y_t = \frac{A(L)}{B(L)} x_t + \varepsilon_t,$$

where $A(L)$ and $B(L)$ are low-order polynomials in the lag operator. Equivalently, we can write

$$B(L)y_t = A(L)x_t + B(L)\varepsilon_t,$$

which emphasizes that the rational distributed lag of x actually brings both lags of x and lags of y into the model. One way or another, it's crucial to allow for lags of y, and we now study such models in greater depth.

16.2 Regressions with Lagged Dependent Variables, and Regressions with *ARMA* Disturbances

There's something missing in distributed lag models of the form

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t.$$

A multivariate model (in this case, a regression model) should relate the current value y to its own past and to the past of x. But as presently written, we've left out the past of y! Even in distributed lag models, we always want to allow for the presence of the usual univariate dynamics. Put differently, the included regressors may not capture all the dynamics in y, which we need to model one way or another. Thus, for example, a preferable model includes lags of the dependent variable,

$$y_t = \beta_0 + \sum_{i=1}^{N_y} \alpha_i y_{t-i} + \sum_{j=1}^{N_x} \delta_j x_{t-j} + \varepsilon_t.$$

This model, a distributed lag regression model with lagged dependent variables, is closely related to, but not exactly the same as, the rational distributed lag model introduced earlier. (Why?) You can think of it as arising by beginning with a univariate autoregressive model for y, and then introducing additional explanatory variables. If the lagged y's don't play a role, as assessed with the usual tests, we can always delete them, but we never want to eliminate from the outset the possibility that lagged dependent variables play a role. Lagged dependent variables absorb residual serial correlation and can *dramatically* enhance forecasting performance.

Alternatively, we can capture own-variable dynamics in distributed-lag regression models by using a distributed-lag regression model with *ARMA* disturbances. Recall that our $ARMA(p, q)$ models are equivalent to regression models, with only a constant regressor, and with $ARMA(p, q)$ disturbances,

$$y_t = \beta_0 + \varepsilon_t$$

$$\varepsilon_t = \frac{\Theta(L)}{\Phi(L)} v_t$$

$$v_t \sim WN(0, \sigma^2).$$

We want to begin with the univariate model as a baseline, and then generalize it to allow for multivariate interaction, resulting in models such as

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t$$

$$\varepsilon_t = \frac{\Theta(L)}{\Phi(L)} v_t$$

$$v_t \sim WN(0, \sigma^2).$$

Regressions with $ARMA$ disturbances make clear that regression (a statistical and econometric tool with a long tradition) and the $ARMA$ model of time-series dynamics (a more recent innovation) are not at all competitors; rather, when used appropriately they can be highly complementary.

It turns out that the distributed-lag regression model with autoregressive disturbances – a great workhorse in econometrics – is a special case of the more general model with lags of both y and x and white noise disturbances. To see this, let's take the simple example of an unconditional (1-step-ahead) regression forecasting model with $AR(1)$ disturbances:

$$y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t$$

$$v_t \ \sim \ WN(0, \ \sigma^2).$$

In lag operator notation, we write the $AR(1)$ regression disturbance as

$$(1-\phi L)\varepsilon_t = v_t,$$

or

$$\varepsilon_t = \frac{1}{(1-\phi L)} v_t.$$

Thus we can rewrite the regression model as

$$y_t = \beta_0 + \beta_1 x_{t-1} + \frac{1}{(1-\phi\, L)}\ v_t.$$

Now multiply both sides by $(1-\phi L)$ to get

$$(1-\phi\, L) y_t = (1-\phi)\beta_0 + \beta_1 (1-\phi\, L) x_{t-1} + v_t,$$

or

$$y_t = \phi y_{t-1} + (1-\phi)\beta_0 + \beta_1 x_{t-1} - \phi\beta_1 x_{t-2} + v_t.$$

Thus a model with one lag of x on the right and $AR(1)$ disturbances is equivalent to a model with y_{t-1}, x_{t-1}, and x_{t-2} on the right-hand side and white noise errors, *subject to the restriction* that the coefficient on the second lag of x_{t-2} is the negative of the product of the coefficients on y_{t-1} *and* x_{t-1}. Thus, distributed lag regressions with lagged dependent variables are more general than distributed lag regressions with dynamic disturbances. In practice, the important thing is to allow for own-variable dynamics *somehow*, in order to account for dynamics in y not explained by the right-hand-side variables. Whether we do so by including lagged dependent variables or by allowing for $ARMA$ disturbances can occasionally be important, but usually it's a comparatively minor issue.

16.3 Vector Autoregressions

A univariate autoregression involves one variable. In a univariate autoregression of order p, we regress a variable on p lags of itself. In contrast, a multivariate autoregression – that is, a vector autoregression, or VAR – involves N variables. In an N-variable vector autoregression of order p, or $VAR(p)$, we estimate N different equations. In each equation, we regress the relevant left-hand-side variable on p lags of itself, *and p lags of every other variable.*[1] Thus the right-hand-side variables are the same in every equation – p lags of every variable.

The key point is that, in contrast to the univariate case, vector autoregressions allow for cross-variable dynamics. Each variable is related not only to its own past, but also to the past of all the other variables in the system. In a two-variable $VAR(1)$, for example, we have two equations, one for each variable (y_1 and y_2) . We write

$$y_{1,t} = \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + \varepsilon_{1,t}$$

$$y_{2,t} = \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + \varepsilon_{2,t}.$$

Each variable depends on one lag of the other variable in addition to one lag of itself; that's one obvious source of multivariate interaction captured by the VAR that may be useful for forecasting. In addition, the disturbances may be correlated, so that when one equation is shocked, the other will typically be shocked as well, which is another type of multivariate interaction that univariate models miss. We summarize the disturbance variance-covariance structure as

$$\begin{aligned}\varepsilon_{1,t} &\sim WN(0,\ \sigma_1^2)\\ \varepsilon_{2,t} &\sim WN(0,\ \sigma_2^2)\end{aligned}$$

[1] Trends, seasonals, and other exogenous variables may also be included, as long as they're all included in every equation.

$$cov(\varepsilon_{1,t},\ \varepsilon_{2,t}) = \sigma_{12}.$$

The innovations *could* be uncorrelated, which occurs when $\sigma_{12} = 0$, but they needn't be.

You might guess that VARs would be hard to estimate. After all, they're fairly complicated models, with potentially many equations and many right-hand-side variables in each equation. In fact, precisely the opposite is true. VARs are very easy to estimate, because we need only run N linear regressions. That's one reason why VARs are so popular – OLS estimation of autoregressive models is simple and stable, in contrast to the numerical estimation required for models with moving-average components.[2] Equation-by-equation OLS estimation also turns out to have very good statistical properties when each equation has the same regressors, as is the case in standard VARs. Otherwise, a more complicated estimation procedure called seemingly unrelated regression, which explicitly accounts for correlation across equation disturbances, would be required to obtain estimates with good statistical properties.[3]

When fitting VARs to data, we use the Schwarz and Akaike criteria, just as in the univariate case. The formulas differ, however, because we're now working with a multivariate system of equations rather than a single equation. To get an AIC or SIC value for a VAR system, we could add up the equation-by-equation AICs or SICs, but unfortunately, doing so is appropriate only if the innovations are uncorrelated across equations, which is a very special and unusual situation. Instead, explicitly multivariate versions of the AIC and SIC – and more advanced formulas – are required that account for cross-equation innovation correlation. It's beyond the scope of this book to derive and present those formulas, because they involve unavoidable use of matrix

[2] Estimation of MA and $ARMA$ models is stable enough in the univariate case but rapidly becomes unwieldy in multivariate situations. Hence multivariate $ARMA$ models are used infrequently in practice, in spite of the potential they hold for providing parsimonious approximations to the Wold representation.

[3] For an exposition of seemingly unrelated regression, see Pindyck and Rubinfeld (1997).

algebra, but fortunately we don't need to. They're pre-programmed in many computer packages, and we interpret the AIC and SIC values computed for VARs of various orders in exactly the same way as in the univariate case: we select that order p such that the AIC or SIC is minimized.

We construct VAR forecasts in a way that precisely parallels the univariate case. We can construct 1-step-ahead point forecasts immediately, because all variables on the right-hand side are lagged by one period. Armed with the 1-step-ahead forecasts, we can construct the 2-step-ahead forecasts, from which we can construct the 3-step-ahead forecasts, and so on in the usual way, following the chain rule of forecasting. We construct interval and density forecasts in ways that also parallel the univariate case. The multivariate nature of VARs makes the derivations more tedious, however, so we bypass them. As always, to construct practical forecasts we replace unknown parameters by estimates.

16.4 Predictive Causality

There's an important statistical notion of causality that's intimately related to forecasting and naturally introduced in the context of VARs. It is based on two key principles: first, cause should occur before effect, and second, a causal series should contain information useful for forecasting that is not available in the other series (including the past history of the variable being forecast). In the unrestricted VARs that we've studied thus far, *everything* causes everything else, because lags of every variable appear on the right of every equation. Cause precedes effect because the right-hand-side variables are lagged, and each variable is useful in forecasting every other variable.

We stress from the outset that the notion of predictive causality contains little if any information about causality in the philosophical sense. Rather, the statement "y_i causes y_j" is just shorthand for the more precise, but longwinded, statement, " y_i contains useful information for predicting y_j (in the linear least squares sense), over and above the past histories of the other variables in the system." To save space, we simply say that y_i causes y_j.

To understand what predictive causality means in the context of a $VAR(p)$, consider the j-th equation of the N-equation system, which has y_j on the left and p lags of each of the N variables on the right. If y_i causes y_j, then at least one of the lags of y_i that appear on the right side of the y_j equation must have a nonzero coefficient.

It's also useful to consider the opposite situation, in which y_i does not cause y_j. In that case, all of the lags of that y_i that appear on the right side of the y_j equation must have zero coefficients.[4] Statistical causality tests are based on this formulation of non-causality. We use an F-test to assess whether all coefficients on lags of y_i are jointly zero.

Note that we've defined non-causality in terms of 1-step-ahead prediction errors. In the bivariate VAR, this implies non-causality in terms of h-step-ahead prediction errors, for all h. (Why?) In higher dimensional cases, things are trickier; 1-step-ahead noncausality does not necessarily imply noncausality at other horizons. For example, variable i may 1-step cause variable j, and variable j may 1-step cause variable k. Thus, variable i 2-step causes variable k, but does not 1-step cause variable k.

Causality tests are often used when building and assessing forecasting models, because they can inform us about those parts of the workings of complicated multivariate models that are particularly relevant for forecasting. Just staring at the coefficients of an estimated VAR (and in complicated systems there are *many* coefficients) rarely yields insights into its workings. Thus we need tools that help us to see through to the practical forecasting properties of the model that concern us. And we often have keen interest in the answers to questions such as "Does y_i contribute toward improving

[4] Note that in such a situation the error variance in forecasting y_j using lags of all variables in the system will be the same as the error variance in forecasting y_j using lags of all variables in the system *except* y_i.

forecasts of y_j?," and "Does y_j contribute toward improving forecasts of y_i?" If the results violate intuition or theory, then we might scrutinize the model more closely. In a situation in which we can't reject a certain noncausality hypothesis, and neither intuition nor theory makes us uncomfortable with it, we might want to *impose* it, by omitting certain lags of certain variables from certain equations.

Various types of causality hypotheses are sometimes entertained. In any equation (the j-th, say), we've already discussed testing the simple noncausality hypothesis that:

(a) No lags of variable i aid in one-step-ahead prediction of variable j.

We can broaden the idea, however. Sometimes we test stronger noncausality hypotheses such as:

(b) No lags of a *set* of other variables aid in one-step-ahead prediction of variable j.

(b) No lags of *any other variables* aid in one-step-ahead prediction of variable j.

All of hypotheses (a), (b) and (c) amount to assertions that various coefficients are zero. Finally, sometimes we test noncausality hypotheses that involve more than one equation, such as:

(b) No variable in a set A causes any variable in a set B, in which case we say that the variables in A are block non-causal for those in B.

This particular noncausality hypothesis corresponds to exclusion restrictions that hold simultaneously in a number of equations. Again, however, standard test procedures are applicable.

16.5 Impulse-Response Functions

The impulse-response function is another device that helps us to learn about the dynamic properties of vector autoregressions of interest to forecasters. We'll introduce it first in the *univariate* context, and then we'll move to VARs. The question of interest is simple and direct: How does a unit innovation to a series affect it, now and in the future? To answer the question, we simply read off the coefficients in the moving average representation of the process.

We're used to normalizing the coefficient on ε_t to unity in moving-average representations, but we don't have to do so; more generally, we can write

$$y_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + ...$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The additional generality introduces ambiguity, however, because we can always multiply and divide every ε_t by an arbitrary constant m, yielding an equivalent model but with different parameters and innovations,

$$y_t = (b_0 m)\left(\frac{1}{m}\varepsilon_t\right) + (b_1 m)\left(\frac{1}{m}\varepsilon_{t-1}\right) + (b_2 m)\left(\frac{1}{m}\varepsilon_{t-2}\right) + ...$$

$$\varepsilon_t \ \sim \ WN(0, \ \sigma^2)$$

or

$$y_t = b'_0\varepsilon'_t + b'_1\varepsilon'_{t-1} + b'_2\varepsilon'_{t-2} + ...$$

$$\varepsilon'_t \sim WN(0, \frac{\sigma^2}{m^2}),$$

where $b'_i = b_i m$ and $\varepsilon'_t = \frac{\varepsilon_t}{m}$.

To remove the ambiguity, we must set a value of m. Typically we set $m = 1$, which yields the standard form of the moving average representation. For impulse-response analysis, however, a different normalization turns out to be particularly convenient; we choose $m = \sigma$, which yields

$$y_t = (b_0\sigma)\left(\frac{1}{\sigma}\varepsilon_t\right) + (b_1\sigma)\left(\frac{1}{\sigma}\varepsilon_{t-1}\right) + (b_2\sigma)\left(\frac{1}{\sigma}\varepsilon_{t-2}\right) + ...$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

or

$$y_t = b'_0\varepsilon'_t + b'_1\varepsilon'_{t-1} + b'_2\varepsilon'_{t-2} + ...$$

$$\varepsilon'_t \sim WN(0, 1),$$

where $b'_i = b_i\sigma$ and $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$. Taking $m = \sigma$ converts shocks to "standard deviation units," because a unit shock to ε'_t corresponds to a one standard deviation shock to ε_t.

To make matters concrete, consider the univariate $AR(1)$ process,

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The standard moving average form is

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

and the equivalent representation in standard deviation units is

$$y_t = b_0\varepsilon'_t + b_1\varepsilon'_{t-1} + b_2\varepsilon'_{t-2} + ...$$

$$\varepsilon'_t \sim WN(0, 1)$$

where $b_i = \phi^i\sigma$ and $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$. The impulse-response function is { b_0, b_1, ... }. The parameter b_0 is the contemporaneous effect of a unit shock to ε'_t, or equivalently a one standard deviation shock to ε_t; as must be the case, then, $b_0 = \sigma$. Note well that b_0 gives the immediate effect of the shock at time t, when it hits. The parameter b_1, which multiplies ε'_{t-1}, gives the effect of the shock one period later, and so on. The full set of impulse-response coefficients, $\{b_0,\ b_1,\ ...\}$, tracks the complete dynamic response of y to the shock.

Now we consider the multivariate case. The idea is the same, but there are more shocks to track. The key question is, "How does a unit shock to ε_i affect y_j, now and in the future, for all the various combinations of i and j?" Consider, for example, the bivariate $VAR(1)$,

$$y_{1t} = \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + \varepsilon_{1t}$$

$$y_{2t} = \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + \varepsilon_{2t}$$

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

$$cov(\varepsilon_1, \varepsilon_2) = \sigma_{12}.$$

The standard moving average representation, obtained by back substitution, is

$$y_{1t} = \varepsilon_{1t} + \phi_{11}\varepsilon_{1,t-1} + \phi_{12}\varepsilon_{2,t-1} + ...$$

$$y_{2t} = \varepsilon_{2t} + \phi_{21}\varepsilon_{1,t-1} + \phi_{22}\varepsilon_{2,t-1} + ...$$

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$

$$cov(\varepsilon_1, \varepsilon_2) = \sigma_{12}.$$

Just as in the univariate case, it proves fruitful to adopt a different normalization of the moving average representation for impulse-response analysis. The multivariate analog of our univariate normalization by σ is called normalization by the Cholesky factor.[5] The resulting VAR moving average representation has a number of useful properties that parallel the univariate case precisely. First, the innovations of the transformed system are in standard deviation units. Second, although the current innovations in the standard representation have unit coefficients, the current innovations in the normalized representation have non-unit coefficients. In fact, the first equation has only one current innovation, ε_{1t}. (The other has a zero coefficient.) The second equation has both current innovations. Thus, the ordering of the variables can matter.[6]

If y_1 is ordered first, the normalized representation is

$$y_{1,t} = b_{11}^0 \varepsilon'_{1,t} + b_{11}^1 \varepsilon'_{1,t-1} + b_{12}^1 \varepsilon'_{2,t-1} + ...$$

$$y_{2,t} = b_{21}^0 \varepsilon'_{1,t} + b_{22}^0 \varepsilon'_{2,t} + b_{21}^1 \varepsilon'_{1,t-1} + b_{22}^1 \varepsilon'_{2,t-1} + ...$$

$$\varepsilon'_{1,t} \sim WN(0,1)$$

$$\varepsilon'_{2,t} \sim WN(0,1)$$

$$cov(\varepsilon'_1, \varepsilon'_2) = 0.$$

Alternatively, if y_2 ordered first, the normalized representation is

$$y_{2,t} = b_{22}^0 \varepsilon'_{2,t} + b_{21}^1 \varepsilon'_{1,t-1} + b_{22}^1 \varepsilon'_{2,t-1} + ...$$

$$y_{1,t} = b_{11}^0 \varepsilon'_{1,t} + b_{12}^0 \varepsilon_{2,t} + b_{11}^1 \varepsilon_{1,t-1} + b_{12}^1 \varepsilon_{2,t-1} + ...$$

$$\varepsilon'_{1,t} \sim WN(0,1)$$

$$\varepsilon'_{2,t} \sim WN(0,1)$$

$$cov(\varepsilon'_1, \varepsilon'_2) = 0.$$

[5] For detailed discussion and derivation of this advanced topic, see Hamilton (1994).

[6] In higher-dimensional VAR's, the equation that's first in the ordering has only one current innovation, ε'_{1t}. The equation that's second has only current innovations ε'_{1t} and ε'_{2t}, the equation that's third has only current innovations ε'_{1t}, ε'_{2t} and ε'_{3t}, and so on.

Finally, the normalization adopted yields a zero covariance between the disturbances of the transformed system. This is crucial, because it lets us perform the experiment of interest – shocking one variable in isolation of the others, which we can do if the innovations are uncorrelated but can't do if they're correlated, as in the original unnormalized representation.

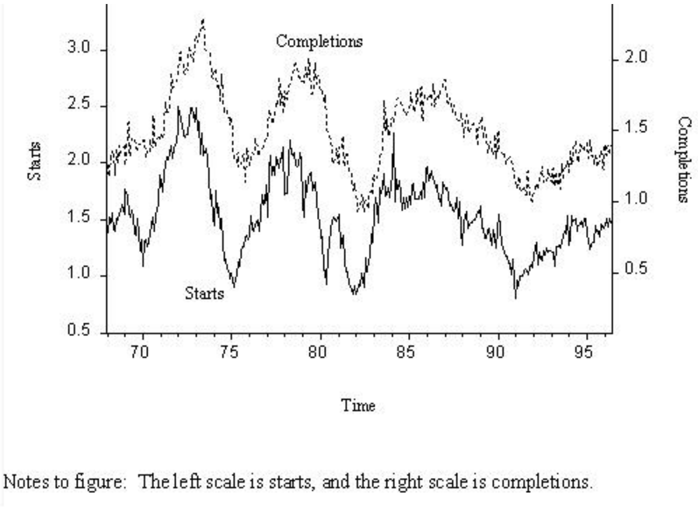
After normalizing the system, for a given ordering, say y_1 first, we compute four sets of impulse-response functions for the bivariate model: response of y_1 to a unit normalized innovation to y_1, $\{\ b^0_{11}, b^1_{11}, b^2_{11}, ...\ \}$, response of y_1 to a unit normalized innovation to y_2, $\{\ b^1_{12}, b^2_{12}, ...\ \}$, response of y_2 to a unit normalized innovation to y_2, $\{\ b^0_{22}, b^1_{22}, b^2_{22}, ...\ \}$, and response of y_2 to a unit normalized innovation to y_1, $\{\ b^0_{21}, b^1_{21}, b^2_{21}, ...\ \}$. Typically we examine the set of impulse-response functions graphically. Often it turns out that impulse-response functions aren't sensitive to ordering, but the only way to be sure is to check.[7]

In practical applications of impulse-response analysis, we simply replace unknown parameters by estimates, which immediately yields point estimates of the impulse-response functions. Getting confidence intervals for impulse-response functions is trickier, however, and adequate procedures are still under development.

16.6 Variance Decompositions

Another way of characterizing the dynamics associated with VARs, closely related to impulse-response functions, is the variance decomposition. Variance decompositions have an immediate link to forecasting – they answer the question, "How much of the h-step-ahead forecast error variance of variable i is explained by innovations to variable j, for $h = 1, 2, ...$" As with impulse-response functions, we typically make a separate graph for every (i, j) pair.

[7] Note well that the issues of normalization and ordering only affect impulse-response analysis; for forecasting we only need the unnormalized model.



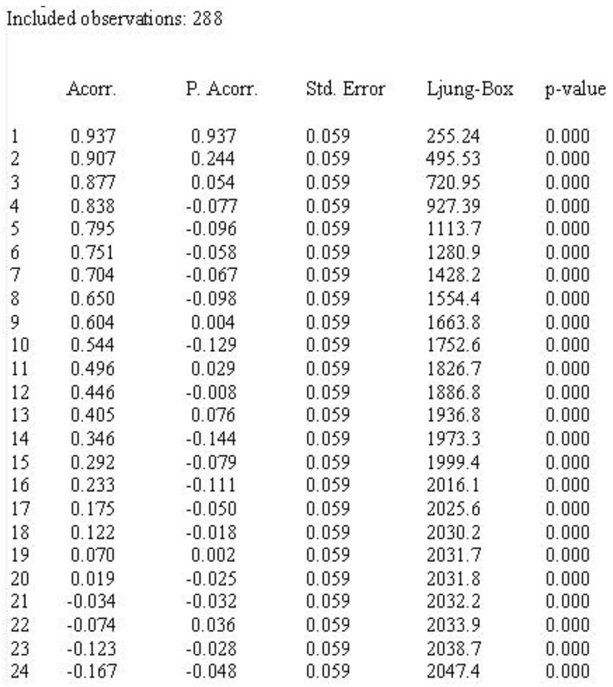
Notes to figure: The left scale is starts, and the right scale is completions.

Figure 16.1: Housing Starts and Completions, 1968 - 1996

Impulse-response functions and the variance decompositions present the same information (although they do so in different ways). For that reason it's not strictly necessary to present both, and impulse-response analysis has gained greater popularity. Hence we offer only this brief discussion of variance decomposition. In the application to housing starts and completions that follows, however, we examine both impulse-response functions and variance decompositions. The two are highly complementary, as with information criteria and correlograms for model selection, and the variance decompositions have a nice forecasting motivation.

16.7 Application: Housing Starts and Completions

We estimate a bivariate VAR for U.S. seasonally-adjusted housing starts and completions, two widely-watched business cycle indicators, 1968.01-1996.06. We use the VAR to produce point extrapolation forecasts. We show housing starts and completions in Figure 16.1. Both are highly cyclical, increasing during business-cycle expansions and decreasing during contractions. Moreover, completions tend to lag behind starts, which makes sense because a house takes time to complete.



Included observations: 288

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.937	0.937	0.059	255.24	0.000
2	0.907	0.244	0.059	495.53	0.000
3	0.877	0.054	0.059	720.95	0.000
4	0.838	-0.077	0.059	927.39	0.000
5	0.795	-0.096	0.059	1113.7	0.000
6	0.751	-0.058	0.059	1280.9	0.000
7	0.704	-0.067	0.059	1428.2	0.000
8	0.650	-0.098	0.059	1554.4	0.000
9	0.604	0.004	0.059	1663.8	0.000
10	0.544	-0.129	0.059	1752.6	0.000
11	0.496	0.029	0.059	1826.7	0.000
12	0.446	-0.008	0.059	1886.8	0.000
13	0.405	0.076	0.059	1936.8	0.000
14	0.346	-0.144	0.059	1973.3	0.000
15	0.292	-0.079	0.059	1999.4	0.000
16	0.233	-0.111	0.059	2016.1	0.000
17	0.175	-0.050	0.059	2025.6	0.000
18	0.122	-0.018	0.059	2030.2	0.000
19	0.070	0.002	0.059	2031.7	0.000
20	0.019	-0.025	0.059	2031.8	0.000
21	-0.034	-0.032	0.059	2032.2	0.000
22	-0.074	0.036	0.059	2033.9	0.000
23	-0.123	-0.028	0.059	2038.7	0.000
24	-0.167	-0.048	0.059	2047.4	0.000

Figure 16.2: Housing Starts Correlogram

We split the data into an estimation sample, 1968.01-1991.12, and a hold-out sample, 1992.01-1996.06 for forecasting. We therefore perform all model specification analysis and estimation, to which we now turn, on the 1968.01-1991.12 data. We show the starts correlogram in Table 16.2 and Figure 16.3. The sample autocorrelation function decays slowly, whereas the sample partial autocorrelation function appears to cut off at displacement 2. The patterns in the sample autocorrelations and partial autocorrelations are highly statistically significant, as evidenced by both the Bartlett standard errors and the Ljung-Box Q-statistics. The completions correlogram, in Table 16.4 and Figure 16.5, behaves similarly.

We've not yet introduced the cross correlation function. There's been no need, because it's not relevant for univariate modeling. It provides important information, however, in the multivariate environments that now concern us. Recall that the autocorrelation function is the correlation between a variable and lags of itself. The cross-correlation function is a natural multivariate analog; it's simply the correlation between a variable and lags of *another*

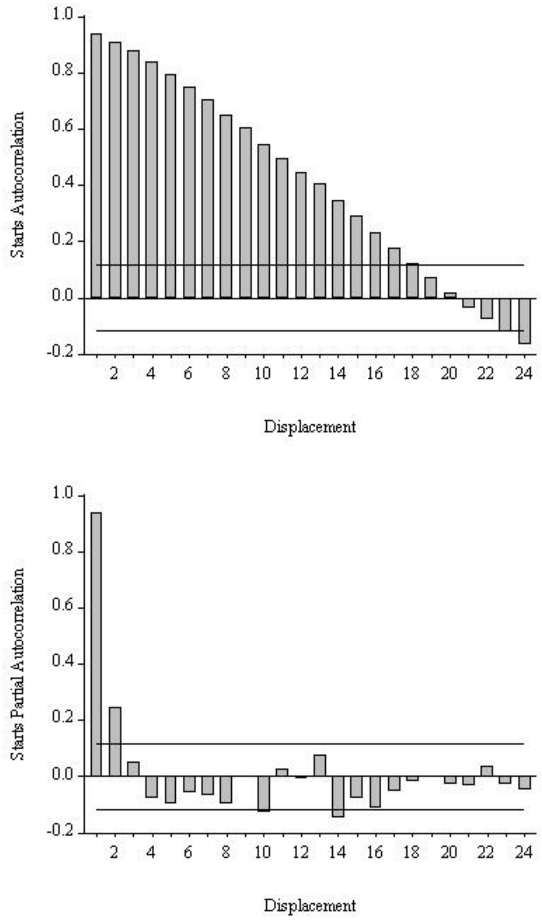
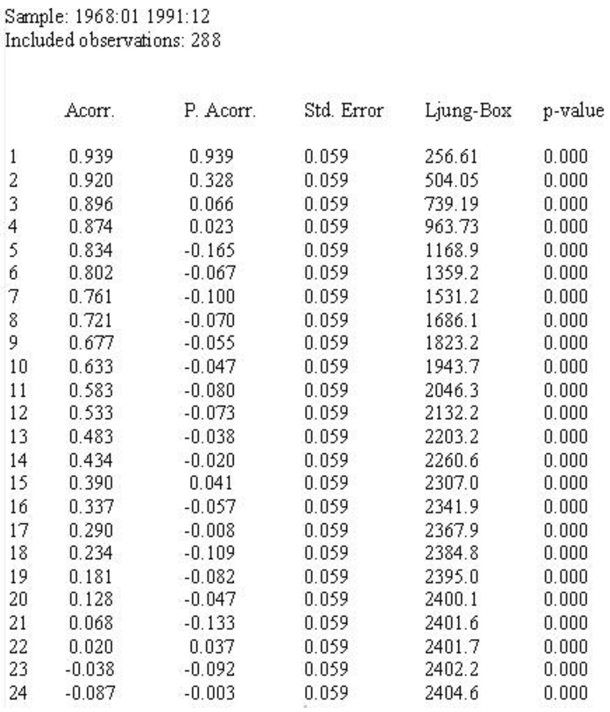


Figure 16.3: Housing Starts Autocorrelations and Partial Autocorrelations

variable. We estimate those correlations using the usual estimator and graph them as a function of displacement along with the Bartlett two- standard-error bands, which apply just as in the univariate case.

The cross-correlation function (Figure 16.6) for housing starts and completions is very revealing. Starts and completions are highly correlated at all displacements, and a clear pattern emerges as well: although the contemporaneous correlation is high (.78), completions are maximally correlated with starts lagged by roughly 6-12 months (around .90). Again, this makes good sense in light of the time it takes to build a house.

Now we proceed to model starts and completions. We need to select the order, p, of our $VAR(p)$. Based on exploration using multivariate versions of SIC and AIC, we adopt a $VAR(4)$.



```
Sample: 1968:01 1991:12
Included observations: 288

     Acorr.   P. Acorr.   Std. Error   Ljung-Box   p-value

1    0.939     0.939      0.059        256.61      0.000
2    0.920     0.328      0.059        504.05      0.000
3    0.896     0.066      0.059        739.19      0.000
4    0.874     0.023      0.059        963.73      0.000
5    0.834    -0.165      0.059        1168.9      0.000
6    0.802    -0.067      0.059        1359.2      0.000
7    0.761    -0.100      0.059        1531.2      0.000
8    0.721    -0.070      0.059        1686.1      0.000
9    0.677    -0.055      0.059        1823.2      0.000
10   0.633    -0.047      0.059        1943.7      0.000
11   0.583    -0.080      0.059        2046.3      0.000
12   0.533    -0.073      0.059        2132.2      0.000
13   0.483    -0.038      0.059        2203.2      0.000
14   0.434    -0.020      0.059        2260.6      0.000
15   0.390     0.041      0.059        2307.0      0.000
16   0.337    -0.057      0.059        2341.9      0.000
17   0.290    -0.008      0.059        2367.9      0.000
18   0.234    -0.109      0.059        2384.8      0.000
19   0.181    -0.082      0.059        2395.0      0.000
20   0.128    -0.047      0.059        2400.1      0.000
21   0.068    -0.133      0.059        2401.6      0.000
22   0.020     0.037      0.059        2401.7      0.000
23  -0.038    -0.092      0.059        2402.2      0.000
24  -0.087    -0.003      0.059        2404.6      0.000
```

Figure 16.4: Housing Completions Correlogram

First consider the starts equation (Table 16.7a), residual plot (Figure 16.7b), and residual correlogram (Table 16.8, Figure 16.9). The explanatory power of the model is good, as judged by the R^2 as well as the plots of actual and fitted values, and the residuals appear white, as judged by the residual sample autocorrelations, partial autocorrelations, and Ljung-Box statistics. Note as well that no lag of completions has a significant effect on starts, which makes sense – we obviously expect starts to cause completions, but not conversely. The completions equation (Table 16.10a), residual plot (Figure 16.10b), and residual correlogram (Table 16.11, Figure 16.12) appear similarly good. Lagged starts, moreover, most definitely have a significant effect on completions.

Table 16.13 shows the results of formal causality tests. The hypothesis that starts don't cause completions is simply that the coefficients on the four lags of starts in the completions equation are all zero. The F-statistic is overwhelmingly significant, which is not surprising in light of the previously-

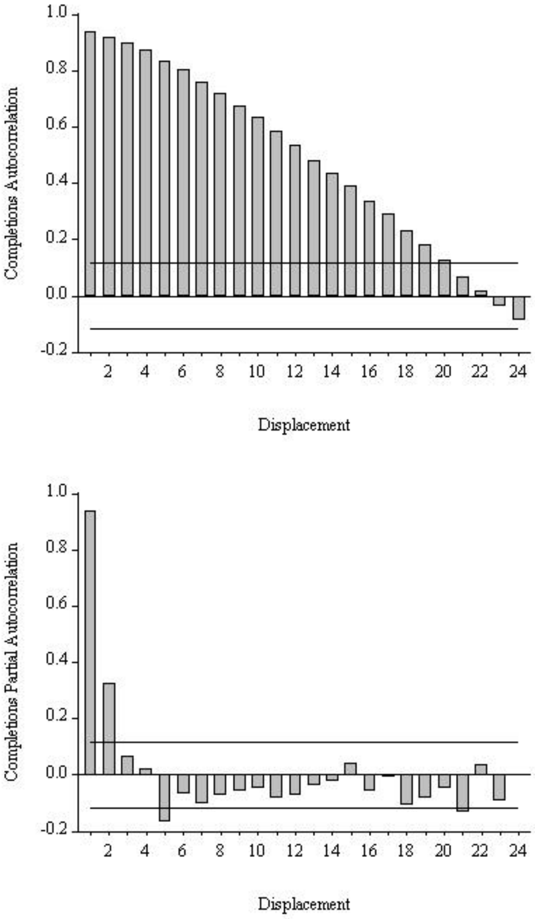
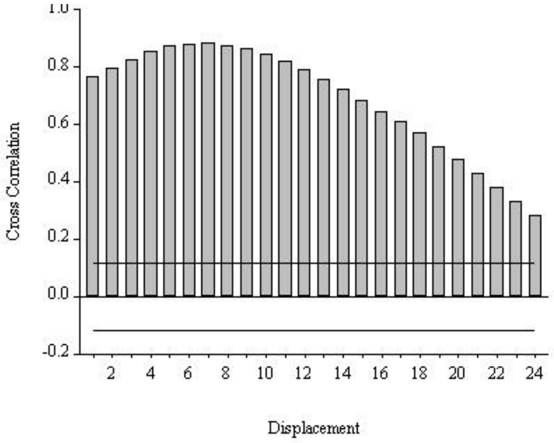
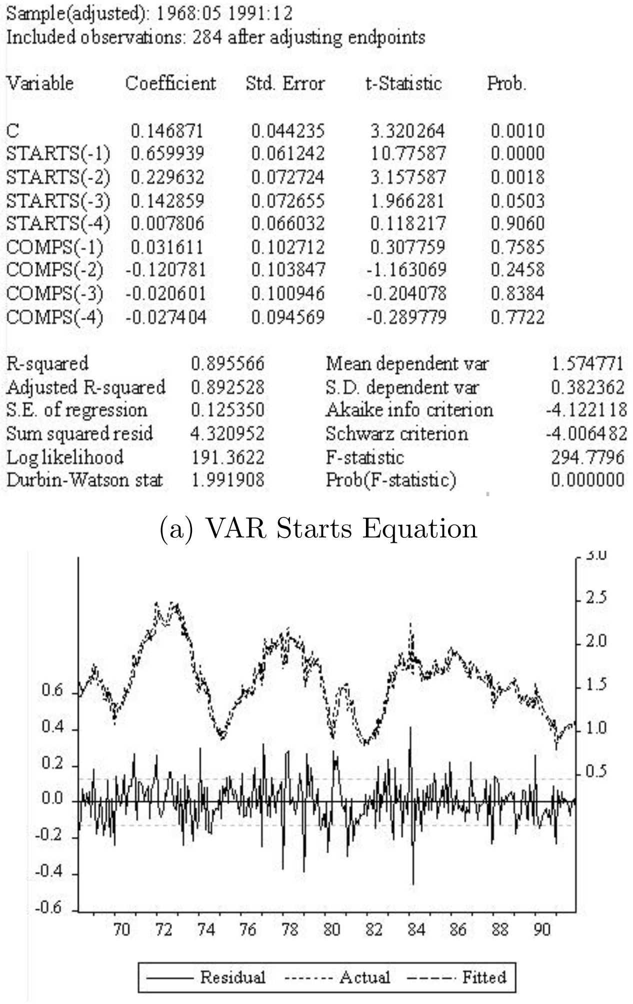


Figure 16.5: Housing Completions Autocorrelations and Partial Autocorrelations



Notes to figure: We graph the sample correlation between completions at time t and starts at time t-i, i = 1, 2, ..., 24.

Figure 16.6: Housing Starts and Completions Sample Cross Correlations



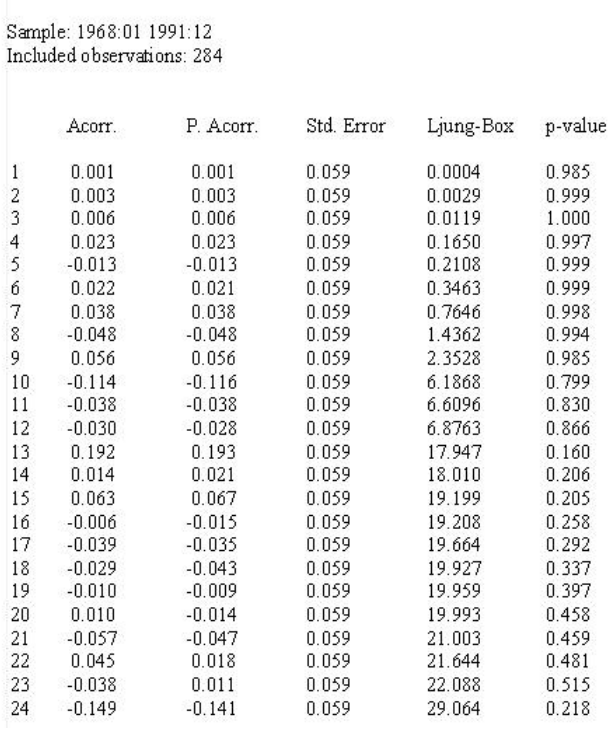
```
Sample(adjusted): 1968:05 1991:12
Included observations: 284 after adjusting endpoints

Variable       Coefficient    Std. Error    t-Statistic    Prob.

C               0.146871      0.044235      3.320264      0.0010
STARTS(-1)      0.659939      0.061242      10.77587      0.0000
STARTS(-2)      0.229632      0.072724      3.157587      0.0018
STARTS(-3)      0.142859      0.072655      1.966281      0.0503
STARTS(-4)      0.007806      0.066032      0.118217      0.9060
COMPS(-1)       0.031611      0.102712      0.307759      0.7585
COMPS(-2)      -0.120781      0.103847     -1.163069      0.2458
COMPS(-3)      -0.020601      0.100946     -0.204078      0.8384
COMPS(-4)      -0.027404      0.094569     -0.289779      0.7722

R-squared             0.895566    Mean dependent var       1.574771
Adjusted R-squared    0.892528    S.D. dependent var       0.382362
S.E. of regression    0.125350    Akaike info criterion   -4.122118
Sum squared resid     4.320952    Schwarz criterion       -4.006482
Log likelihood        191.3622    F-statistic              294.7796
Durbin-Watson stat    1.991908    Prob(F-statistic)        0.000000
```

(a) VAR Starts Equation

(b) VAR Starts Equation - Residual Plot

Figure 16.7: VAR Starts Model

noticed highly-significant t-statistics. Thus we reject noncausality from starts to completions at any reasonable level. Perhaps more surprising, we also reject noncausality from completions to starts at roughly the 5% level. Thus the causality appears bi-directional, in which case we say there is feedback.



```
Sample: 1968:01 1991:12
Included observations: 284

      Acorr.      P. Acorr.    Std. Error   Ljung-Box   p-value

1     0.001       0.001        0.059        0.0004      0.985
2     0.003       0.003        0.059        0.0029      0.999
3     0.006       0.006        0.059        0.0119      1.000
4     0.023       0.023        0.059        0.1650      0.997
5     -0.013      -0.013       0.059        0.2108      0.999
6     0.022       0.021        0.059        0.3463      0.999
7     0.038       0.038        0.059        0.7646      0.998
8     -0.048      -0.048       0.059        1.4362      0.994
9     0.056       0.056        0.059        2.3528      0.985
10    -0.114      -0.116       0.059        6.1868      0.799
11    -0.038      -0.038       0.059        6.6096      0.830
12    -0.030      -0.028       0.059        6.8763      0.866
13    0.192       0.193        0.059        17.947      0.160
14    0.014       0.021        0.059        18.010      0.206
15    0.063       0.067        0.059        19.199      0.205
16    -0.006      -0.015       0.059        19.208      0.258
17    -0.039      -0.035       0.059        19.664      0.292
18    -0.029      -0.043       0.059        19.927      0.337
19    -0.010      -0.009       0.059        19.959      0.397
20    0.010       -0.014       0.059        19.993      0.458
21    -0.057      -0.047       0.059        21.003      0.459
22    0.045       0.018        0.059        21.644      0.481
23    -0.038      0.011        0.059        22.088      0.515
24    -0.149      -0.141       0.059        29.064      0.218
```

Figure 16.8: VAR Starts Residual Correlogram

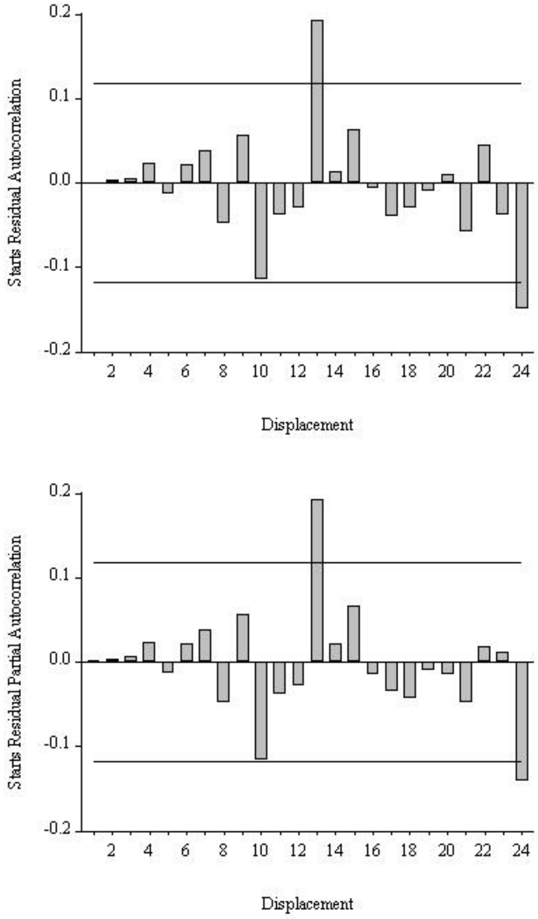
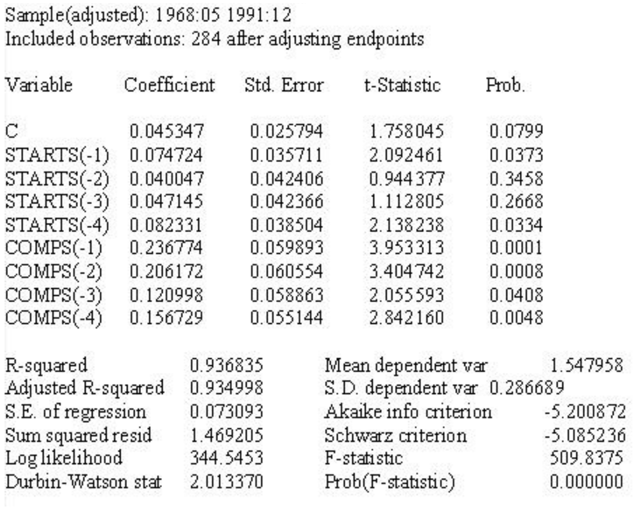
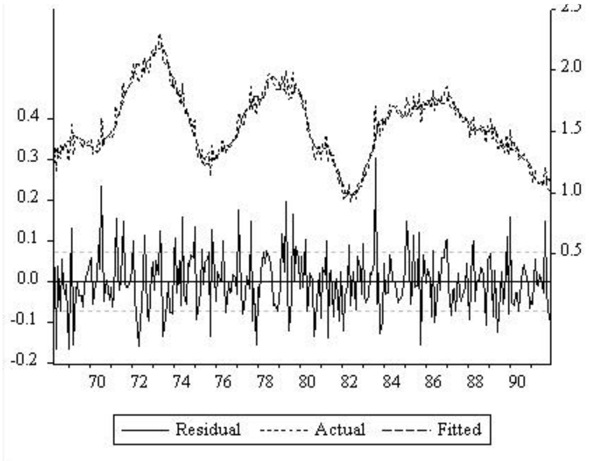


Figure 16.9: VAR Starts Equation - Sample Autocorrelation and Partial Autocorrelation



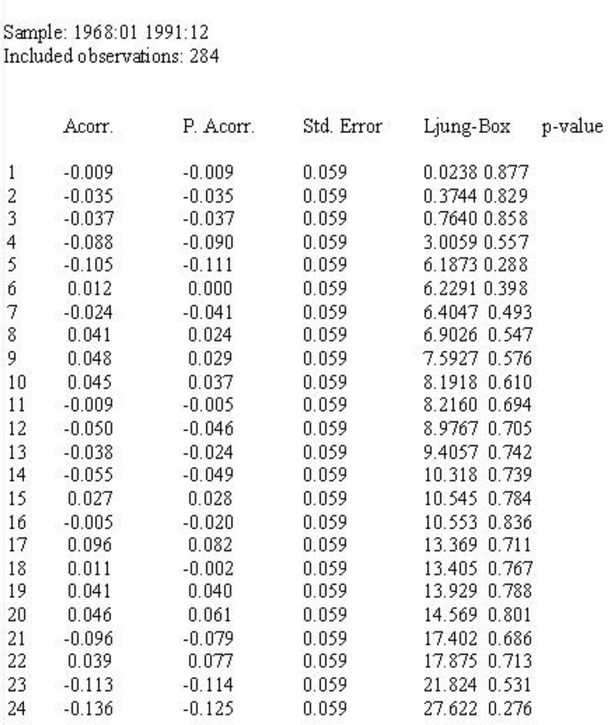
```
Sample(adjusted): 1968:05 1991:12
Included observations: 284 after adjusting endpoints

Variable      Coefficient   Std. Error   t-Statistic   Prob.

C              0.045347      0.025794     1.758045     0.0799
STARTS(-1)     0.074724      0.035711     2.092461     0.0373
STARTS(-2)     0.040047      0.042406     0.944377     0.3458
STARTS(-3)     0.047145      0.042366     1.112805     0.2668
STARTS(-4)     0.082331      0.038504     2.138238     0.0334
COMPS(-1)      0.236774      0.059893     3.953313     0.0001
COMPS(-2)      0.206172      0.060554     3.404742     0.0008
COMPS(-3)      0.120998      0.058863     2.055593     0.0408
COMPS(-4)      0.156729      0.055144     2.842160     0.0048

R-squared             0.936835    Mean dependent var       1.547958
Adjusted R-squared    0.934998    S.D. dependent var  0.286689
S.E. of regression    0.073093    Akaike info criterion   -5.200872
Sum squared resid     1.469205    Schwarz criterion       -5.085236
Log likelihood        344.5453    F-statistic              509.8375
Durbin-Watson stat    2.013370    Prob(F-statistic)        0.000000
```

(a) VAR Completions Equation



(b) VAR Completions Equation - Residual Plot

Figure 16.10: VAR Completions Model



```
Sample: 1968:01 1991:12
Included observations: 284

     Acorr.      P. Acorr.    Std. Error    Ljung-Box    p-value

1    -0.009      -0.009       0.059         0.0238 0.877
2    -0.035      -0.035       0.059         0.3744 0.829
3    -0.037      -0.037       0.059         0.7640 0.858
4    -0.088      -0.090       0.059         3.0059 0.557
5    -0.105      -0.111       0.059         6.1873 0.288
6     0.012       0.000       0.059         6.2291 0.398
7    -0.024      -0.041       0.059         6.4047  0.493
8     0.041       0.024       0.059         6.9026  0.547
9     0.048       0.029       0.059         7.5927  0.576
10    0.045       0.037       0.059         8.1918  0.610
11   -0.009      -0.005       0.059         8.2160  0.694
12   -0.050      -0.046       0.059         8.9767  0.705
13   -0.038      -0.024       0.059         9.4057  0.742
14   -0.055      -0.049       0.059         10.318  0.739
15    0.027       0.028       0.059         10.545  0.784
16   -0.005      -0.020       0.059         10.553  0.836
17    0.096       0.082       0.059         13.369  0.711
18    0.011      -0.002       0.059         13.405  0.767
19    0.041       0.040       0.059         13.929  0.788
20    0.046       0.061       0.059         14.569  0.801
21   -0.096      -0.079       0.059         17.402  0.686
22    0.039       0.077       0.059         17.875  0.713
23   -0.113      -0.114       0.059         21.824  0.531
24   -0.136      -0.125       0.059         27.622  0.276
```

Figure 16.11: VAR Completions Residual Correlogram

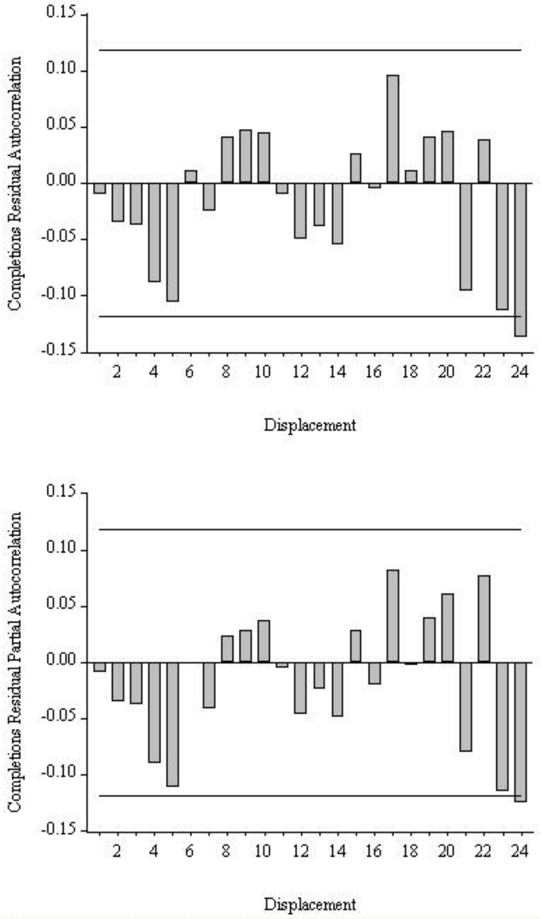
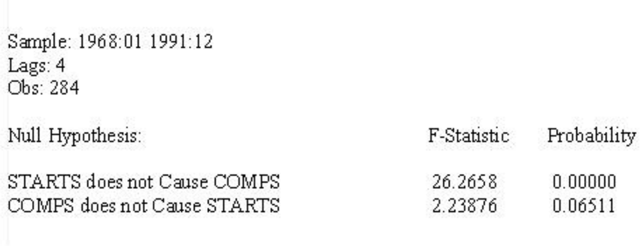


Figure 16.12: VAR Completions Equation - Sample Autocorrelation and Partial Autocorrelation



```
Sample: 1968:01 1991:12
Lags: 4
Obs: 284

Null Hypothesis:                        F-Statistic    Probability

STARTS does not Cause COMPS              26.2658        0.00000
COMPS does not Cause STARTS              2.23876        0.06511
```

Figure 16.13: Housing Starts and Completions - Causality Tests

In order to get a feel for the dynamics of the estimated VAR before producing forecasts, we compute impulse-response functions and variance decompositions. We present results for starts first in the ordering, so that a current innovation to starts affects only current starts, but the results are robust to reversal of the ordering.

In Figure 16.14, we display the impulse-response functions. First let's consider the own-variable impulse responses, that is, the effects of a starts innovation on subsequent starts or a completions innovation on subsequent completions; the effects are similar. In each case, the impulse response is large and decays in a slow, approximately monotonic fashion. In contrast, the cross-variable impulse responses are very different. An innovation to starts produces no movement in completions at first, but the effect gradually builds and becomes large, peaking at about fourteen months. (It takes time to build houses.) An innovation to completions, however, produces little movement in starts at any time. Figure 16.15 shows the variance decompositions. The fraction of the error variance in forecasting starts due to innovations in starts is close to 100 percent at all horizons. In contrast, the fraction of the error variance in forecasting completions due to innovations in starts is near zero at short horizons, but it rises steadily and is near 100 percent at long horizons, again reflecting time-to-build effects.

Finally, we construct forecasts for the out-of-sample period, 1992.01-1996.06. The starts forecast appears in Figure 16.16. Starts begin their recovery before 1992.01, and the VAR projects continuation of the recovery. The VAR fore-

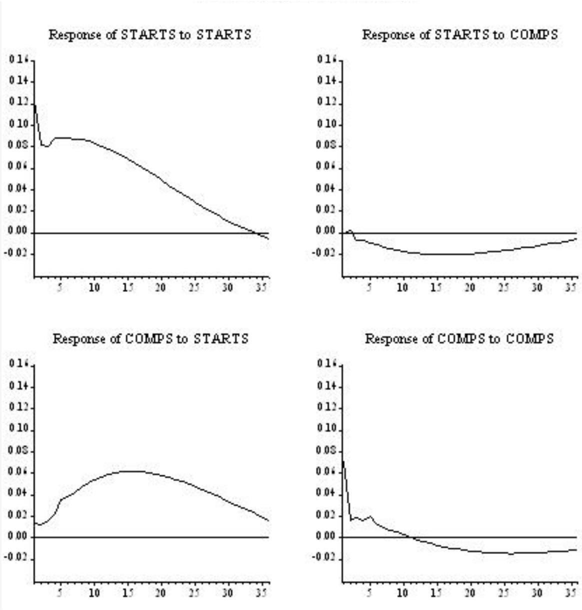


Figure 16.14: Housing Starts and Completions - VAR Impulse Response Functions. Response is to 1 SD innovation.

casts captures the general pattern quite well, but it forecasts quicker mean reversion than actually occurs, as is clear when comparing the forecast and realization in Figure 16.17. The figure also makes clear that the recovery of housing starts from the recession of 1990 was slower than the previous recoveries in the sample, which naturally makes for difficult forecasting. The completions forecast suffers the same fate, as shown in Figures 16.18 and 16.19. Interestingly, however, completions had not yet turned by 1991.12, but the forecast nevertheless correctly predicts the turning point. (Why?)

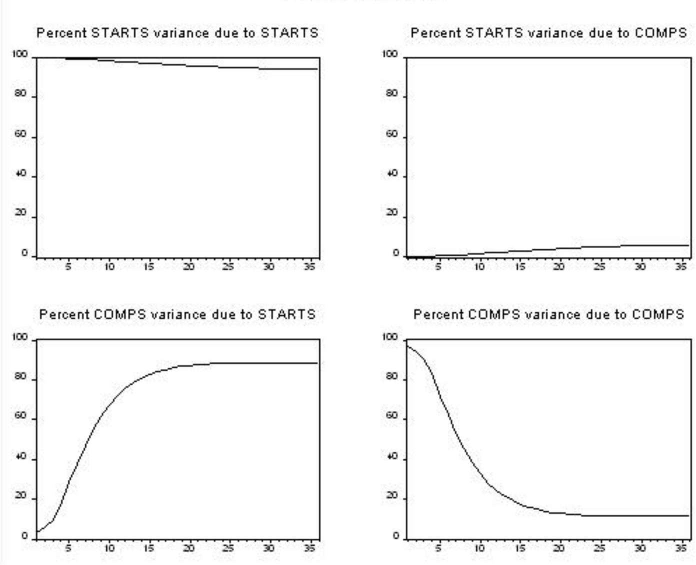


Figure 16.15: Housing Starts and Completions - VAR Variance Decompositions

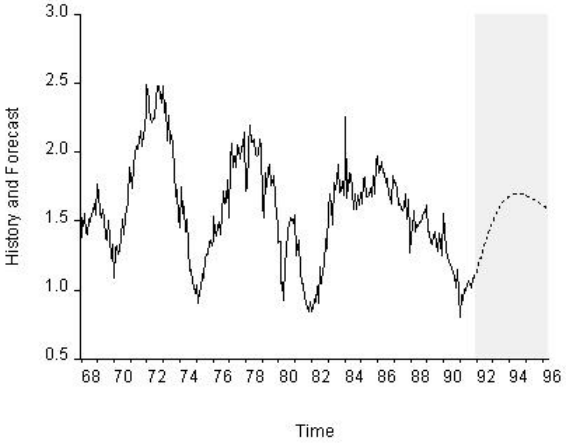


Figure 16.16: Housing Starts Forecast

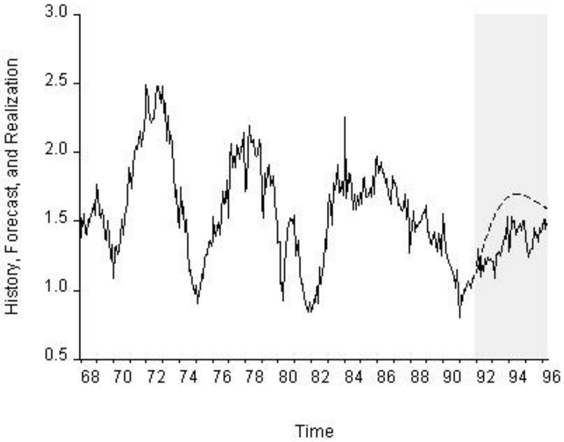


Figure 16.17: Housing Starts Forecast and Realization

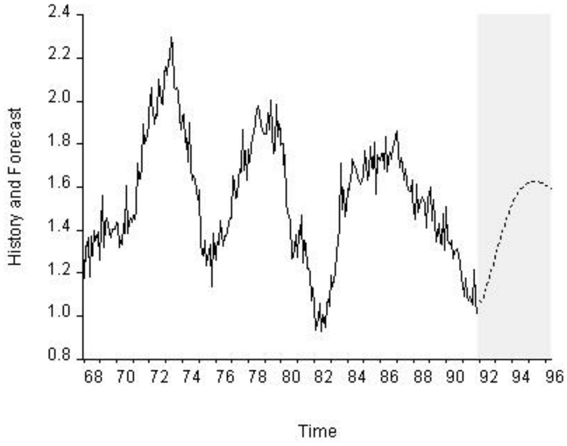


Figure 16.18: Housing Completions Forecast

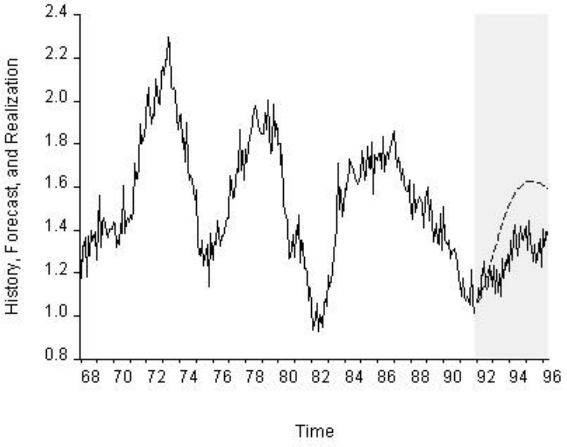


Figure 16.19: Housing Completions Forecast and Realization

16.8 Exercises, Problems and Complements

1. Housing starts and completions, continued.

   Our VAR analysis of housing starts and completions, as always, involved many judgment calls. Using the starts and completions data, assess the adequacy of our models and forecasts. Among other things, you may want to consider the following questions:

   a. Should we allow for a trend in the forecasting model?

   b. How do the results change if, in light of the results of the causality tests, we exclude lags of completions from the starts equation, re-estimate by seemingly-unrelated regression, and forecast?

   c. Are the VAR forecasts of starts and completions more accurate than univariate forecasts?

2. Forecasting crop yields.

   Consider the following dilemma in agricultural crop yield forecasting:

   The possibility of forecasting crop yields several years in advance would, of course, be of great value in the planning of agricultural production. However, the success of long-range crop forecasts is contingent not only on our knowledge of the weather factors determining yield, but also on our ability to predict the weather. Despite an abundant literature in this field, no firm basis for reliable long-range weather forecasts has yet been found. (Sanderson, 1953, p. 3)

   a. How is the situation related to our concerns in this chapter, and specifically, to the issue of conditional vs. unconditional forecasting?

   b. What variables other than weather might be useful for predicting crop yield?

   c. How would you suggest that the forecaster should proceed?

3. Econometrics, time series analysis, and forecasting.

   As recently as the early 1970s, time series analysis was mostly univariate and made little use of economic theory. Econometrics, in contrast, stressed the cross-variable dynamics associated with economic theory, with equations estimated using multiple regression. Econometrics, moreover, made use of simultaneous systems of such equations, requiring complicated estimation methods. Thus the econometric and time series approaches to forecasting were very different.[8]

   As Klein (1981) notes, however, the complicated econometric system estimation methods had little payoff for practical forecasting and were therefore largely abandoned, whereas the rational distributed lag patterns associated with time-series models led to large improvements in practical forecast accuracy.[9] Thus, in more recent times, the distinction between econometrics and time series analysis has largely vanished, with the union incorporating the best of both. In many respects the VAR is a modern embodiment of both econometric and time-series traditions. VARs use economic considerations to determine which variables to include and which (if any) restrictions should be imposed, allow for rich multivariate dynamics, typically require only simple estimation techniques, and are explicit forecasting models.

4. Business cycle analysis and forecasting: expansions, contractions, turning points, and leading indicators[10].

   The use of anticipatory data is linked to business cycle analysis in general, and leading indicators in particular. During the first half of this

[8] Klein and Young (1980) and Klein (1983) provide good discussions of the traditional econometric simultaneous equations paradigm, as well as the link between structural simultaneous equations models and reduced-form time series models. Wallis (1995) provides a good summary of modern large-scale macroeconometric modeling and forecasting, and Pagan and Robertson (2002) provide an intriguing discussion of the variety of macroeconomic forecasting approaches currently employed in central banks around the world.

[9] For an acerbic assessment circa the mid-1970s, see Jenkins (1979).

[10] This complement draws in part upon Diebold and Rudebusch (1996).

century, much research was devoted to obtaining an empirical characterization of the business cycle. The most prominent example of this work was Burns and Mitchell (1946), whose summary empirical definition was:

Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle. (p. 3)

The comovement among individual economic variables was a key feature of Burns and Mitchell's definition of business cycles. Indeed, the comovement among series, taking into account possible leads and lags in timing, was the centerpiece of Burns and Mitchell's methodology. In their analysis, Burns and Mitchell considered the historical concordance of hundreds of series, including those measuring commodity output, income, prices, interest rates, banking transactions, and transportation services, and they classified series as leading, lagging or coincident. One way to define a leading indicator is to say that a series x is a leading indicator for a series y if x causes y in the predictive sense. According to that definition, for example, our analysis of housing starts and completions indicates that starts are a leading indicator for completions.

Leading indicators have the potential to be used in forecasting equations in the same way as anticipatory variables. Inclusion of a leading indicator, appropriately lagged, can improve forecasts. Zellner and Hong (1989) and Zellner, Hong and Min (1991), for example, make good use of that idea in their ARLI (autoregressive leading-indicator) models for forecasting aggregate output growth. In those models, Zellner *et al.* build forecasting models by regressing output on lagged output and lagged leading indicators; they also use shrinkage techniques to coax the forecasted growth rates toward the international average, which improves forecast performance.

Burns and Mitchell used the clusters of turning points in individual series to determine the monthly dates of the turning points in the overall business cycle, and to construct composite indexes of leading, coincident, and lagging indicators. Such indexes have been produced by the National Bureau of Economic Research (a think tank in Cambridge, Mass.), the Department of Commerce (a U.S. government agency in Washington, DC), and the Conference Board (a business membership organization based in New York).[11] Composite indexes of leading indicators are often used to gauge likely future economic developments, but their usefulness is by no means uncontroversial and remains the subject of ongoing research. For example, leading indexes apparently cause aggregate output in analyses of ex post historical data (Auerbach, 1982), but they appear much less useful in real-time forecasting, which is what's relevant (Diebold and Rudebusch, 1991).

5. Spurious regression.

   Consider two variables y and x, both of which are highly serially correlated, as are most series in business, finance and economics. Suppose in addition that y and x are completely unrelated, but that we don't know they're unrelated, and we regress y on x using ordinary least squares.

   a. If the usual regression diagnostics (e.g., R^2, t-statistics, F-statistic) were reliable, we'd expect to see small values of all of them. Why?

   b. In fact the opposite occurs; we tend to see large R^2, t-, and F-statistics, and *a very low Durbin-Watson statistic*. Why the low

[11] The indexes build on very early work, such as the Harvard "Index of General Business Conditions." For a fascinating discussion of the early work, see Hardy (1923), Chapter 7.

Durbin-Watson? Why, given the low Durbin-Watson, might you *expect* misleading R^2, t-, and F-statistics?

c. This situation, in which highly persistent series that are in fact unrelated nevertheless appear highly related, is called spurious regression. Study of the phenomenon dates to the early twentieth century, and a key study by Granger and Newbold (1974) drove home the prevalence and potential severity of the problem. How might you insure yourself against the spurious regression problem? (Hint: Consider allowing for lagged dependent variables, or dynamics in the regression disturbances, as we've advocated repeatedly.)

6. Comparative forecasting performance of VARs and univariate models.

   Using the housing starts and completions data on the book's website, compare the forecasting performance of the VAR used in this chapter to that of the obvious competitor: univariate autoregressions. Use the same in-sample and out-of-sample periods as in the chapter. Why might the forecasting performance of the VAR and univariate methods differ? Why might you expect the VAR completions forecast to outperform the univariate autoregression, but the VAR starts forecast to be no better than the univariate autoregression? Do your results support your conjectures?

7. VARs as Reduced Forms of Simultaneous Equations Models.

   VARs look restrictive in that only *lagged* values appear on the right. That is, the LHS variables are not contemporaneously affected by other variables – instead they are contemporaneously affected only by shocks. That appearance is deceptive, however, as simultaneous equations systems have VAR reduced forms. Consider, for example, the simultaneous system

$$(A_0 + A_1 L + ... + A_p L^p) y_t = v_t$$

$$v_t \sim iid(0, \Omega).$$

Mutiplying through by A_0^{-1} yields

$$(I + A_0^{-1}A_1L + ... + A_0^{-1}A_pL^p)y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, A_0^{-1}\Omega A_0^{-1\prime})$$

or

$$(I + \Phi_1L + ... + \Phi_pL^p)y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \Sigma)$$

$$\Sigma = A_0^{-1}\Omega A_0^{-1\prime},$$

which is a standard VAR. The VAR structure, moreover, is needed for forecasting, as everything on the RHS is lagged by at least one period, making Wold's chain rule immediately applicable.

8. Transfer Function Models.

We saw that distributed lag regressions with lagged dependent variables are more general than distributed lag regressions with dynamic disturbances. Transfer function models are more general still, and include both as special cases.[12] The basic idea is to exploit the power and parsimony of rational distributed lags in modeling both own-variable and cross-variable dynamics. Imagine beginning with a univariate $ARMA$ model,

$$y_t = \frac{C(L)}{D(L)}\varepsilon_t,$$

which captures own-variable dynamics using a rational distributed lag. Now extend the model to capture cross-variable dynamics using a rational distributed lag of the other variable, which yields the general transfer

[12] Table 1 displays a variety of important forecasting models, all of which are special cases of the transfer function model.

function model,

$$y_t = \frac{A(L)}{B(L)} x_t + \frac{C(L)}{D(L)} \varepsilon_t.$$

Distributed lag regression with lagged dependent variables is a potentially restrictive special case, which emerges when $C(L) = 1$ and $B(L) = D(L)$. (Verify this for yourself.) Distributed lag regression with $ARMA$ disturbances is also a special case, which emerges when $B(L) = 1$. (Verify this too.) In practice, the important thing is to allow for own-variable dynamics *somehow*, in order to account for dynamics in y not explained by the RHS variables. Whether we do so by including lagged dependent variables, or by allowing for $ARMA$ disturbances, or by estimating general transfer function models, can occasionally be important, but usually it's a comparatively minor issue.

9. Cholesky-Factor Identified VARs in Matrix Notation.

10. Inflation Forecasting via "Structural" Phillps-Curve Models vs. Time-Series Models.

    The literature started with Atkinson and Ohanian ****. The basic result is that Phillips curve information doesn't improve on univariate time series, which is interesting. Also interesting is thinking about why. For example, the univariate time series used is often $IMA(0, 1, 1)$ (i.e., exponential smoothing, or local level), which Hendry, Clements and others have argued is robust to shifts. Maybe that's why exponential smoothing is still so powerful after all these years.

11. Multivariate point forecast evaluation.

    All univariate absolute standards continue to hold, appropriately interpreted.

    – Zero-mean error vector.

    – 1-step-ahead errors are vector white noise.

– h-step-ahead errors are at most vector $MA(h-1)$.

– h-step-ahead error covariance matrices are non-decreasing in h. That is, $\Sigma_h - \Sigma_{h-1}$ is p.s.d. for all $h > 1$.

– The error vector is orthogonal to all available information.

Relative standards, however, need more thinking, as per Christoffersen and Diebold (1998) and Primiceri, Giannone and Lenza (2014). $trace(MSE)$, $e'Ie$ is not necessarily adequate, and neither is $e'De$ for diagonal d; rather, we generally want $e'\Sigma e$, so as to reflect preferences regarding multivariate interactions.

12. Multivariate density forecast evaluation

The principle that governs the univariate techniques in this paper extends to the multivariate case, as shown in Diebold, Hahn and Tay (1998). Suppose that the variable of interest y is now an $(N \times 1)$ vector, and that we have on hand m multivariate forecasts and their corresponding multivariate realizations. Further suppose that we are able to decompose each period's forecasts into their conditionals, i.e., for each period's forecasts we can write

$$p(y_{1t}, y_{2t}, ..., y_{Nt}|\Phi_{t-1}) \;=\; p(y_{Nt}|y_{N-1,t}, ..., y_{1t}, \Phi_{t-1})...p(y_{2t}|y_{1t}, \Phi_{t-1})p(y_{1t}|\Phi_{t-1}),$$

where Φ_{t-1} now refers to the past history of $(y_{1t}, y_{2t}, ..., y_{Nt})$. Then for each period we can transform each element of the multivariate observation $(y_{1t}, y_{2t}, ..., y_{Nt})$ by its corresponding conditional distribution. This procedure will produce a set of N z series that will be iid $U(0, 1)$ individually, and also when taken as a whole, if the multivariate density forecasts are correct. Note that we will have $N!$ sets of z series, depending on how the joint density forecasts are decomposed, giving us a wealth of information with which to evaluate the forecasts. In addition, the univariate formula for the adjustment of forecasts, discussed above, can be applied to each individual conditional, yielding

$$\begin{aligned} f(y_{1t}, y_{2t}, ..., y_{Nt}|\Phi_{t-1}) = & \prod_{i=1}^{N} [p(y_{it}|y_{i-1,t}, ..., y_{1t}, \Phi_{t-1}) q(P(y_{it}|y_{i-1,t}, ..., y_{1t}, \Phi_{t-1}))] \\ = & \ p(y_{1t}, y_{2t}, ..., y_{Nt}|\Phi_{t-1}) q(z_{1t}, z_{2t}, ..., z_{Nt}|\Phi_{t-1}) \ . \end{aligned}$$

16.9 Notes

Some software, such as Eviews, automatically accounts for parameter uncertainty when forming conditional regression forecast intervals by using variants of the techniques we introduced in Section ***. Similar but advanced techniques are sometimes used to produce unconditional forecast intervals for dynamic models, such as autoregressions (see Lütkepohl, 1991), but bootstrap simulation techniques are becoming increasingly popular (Efron and Tibshirani, 1993).

Chatfield (1993) argues that innovation uncertainty and parameter estimation uncertainty are likely of minor importance compared to specification uncertainty. We rarely acknowledge specification uncertainty, because we don't know how to quantify "what we don't know we don't know." Quantifying it is a major challenge for future research, and useful recent work in that direction includes Chatfield (1995).

The idea that regression models with serially correlated disturbances are more restrictive than other sorts of transfer function models has a long history in econometrics and engineering and is highlighted in a memorably-titled paper, "Serial Correlation as a Convenient Simplification, not a Nuisance," by Hendry and Mizon (1978). Engineers have scolded econometricians for not using more general transfer function models, as for example in Jenkins (1979). But the fact is, as we've seen repeatedly, that generality for generality's sake in business and economic forecasting is not necessarily helpful, and can be positively harmful. The shrinkage principle asserts that the imposition of restrictions – even false restrictions – can be helpful in forecasting.

Sims (1980) is an influential paper arguing the virtues of VARs. The idea of predictive causality and associated tests in VARs is due to Granger (1969) and Sims (1972), who build on earlier work by the mathematician Norbert Weiner. Lütkepohl (1991) is a good reference on VAR analysis and forecasting.

Gershenfeld and Weigend (1993) provide a perspective on time series forecasting from the computer-science/engineering/nonlinear/neural-net perspective, and Swanson and White (1995) compare and contrast a variety of linear and nonlinear forecasting methods.

Some slides that might be usefully incorporated:

Univariate $AR(p)$:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t$$

$$y_t = \phi_1 L y_t + ... + \phi_p L^p y_t + \varepsilon_t$$

$$(I - \phi_1 L - ... - \phi_p L^p) y_t = \varepsilon_t$$

$$\phi(L) y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

But what if we have more than 1 "y" variable?

Cross-variable interactions? Leads? Lags? Causality?

N-Variable $VAR(p)$

$$y_{1t} = \phi_{11}^1 y_{1,t-1} + ... + \phi_{1N}^1 y_{N,t-1} + ... + \phi_{11}^p y_{1,t-p} + ... + \phi_{1N}^p y_{N,t-p} + \varepsilon_{1t}$$

$\vdots$

$$y_{Nt} = \phi_{N1}^1 y_{1,t-1} + ... + \phi_{NN}^1 y_{N,t-1} + ... + \phi_{N1}^p y_{1,t-p} + ... + \phi_{NN}^p y_{N,t-p} + \varepsilon_{Nt}$$

$$\begin{pmatrix} y_{1t} \\ \vdots \\ y_{Nt} \end{pmatrix} = \begin{pmatrix} \phi_{11}^1 & ... & \phi_{1N}^1 \\ \vdots & & \vdots \\ \phi_{N1}^1 & ... & \phi_{NN}^1 \end{pmatrix} \begin{pmatrix} y_{1,t-1} \\ \vdots \\ y_{N,t-1} \end{pmatrix} + ... + \begin{pmatrix} \phi_{11}^p & ... & \phi_{1N}^p \\ \vdots & & \vdots \\ \phi_{N1}^p & ... & \phi_{NN}^p \end{pmatrix} \begin{pmatrix} y_{1,t-p} \\ \vdots \\ y_{N,t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{Nt} \end{pmatrix}$$

$$y_t = \Phi_1 y_{t-1} + ... + \Phi_p y_{t-p} + \varepsilon_t$$

$$y_t = \Phi_1 L y_t + ... + \Phi_p L^p y_t + \varepsilon_t$$

$$(I - \Phi_1 L - ... - \Phi_p L^p) y_t = \varepsilon_t$$

$$\Phi(L)y_t = \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \Sigma)$$

Estimation and Selection

Estimation: Equation-by-equation OLS

Selection: AIC, SIC

$$AIC = \frac{-2lnL}{T} + \frac{2K}{T}$$

$$SIC = \frac{-2lnL}{T} + \frac{KlnT}{T}$$

The Cross-Correlation Function

Recall the univariate autocorrelation function:

$$\rho_y(\tau) = corr(y_t, y_{t-\tau})$$

In multivariate environments we also have
the cross-correlation function:

$$\rho_{yx}(\tau) = corr(y_t, x_{t-\tau})$$

Granger-Sims Causality

Bivariate case:

y_i Granger-Sims causes y_j if
y_i has predictive content for y_j,
over and above the past history of y_j.

Testing:

Are lags of y_i significant in the y_j equation?

Impulse-Response Functions in $AR(1)$ Case

$$y_t = \phi y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim iid(0, \sigma^2)$$

$$\implies \quad y_t = B(L)\varepsilon_t = \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + ...$$

$$= \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + ...$$

IRF is $\{1,\ \phi,\ \phi^2, ...\}$ "dynamic response to a unit shock in ε"

Alternatively write $\varepsilon_t = \sigma v_t,\ v_t \sim iid(0,1)$

$$\implies \quad y_t = \sigma v_t + (\phi\sigma)v_{t-1} + (\phi^2\sigma)v_{t-2} + ...$$

IRF is $\{\sigma,\ \phi\sigma,\ \phi^2\sigma, ...\}$ "dynamic response to a one-σ shock in ε"

Impulse-Response Functions in $VAR(p) Case$

$$y_t = \Phi y_{t-1} + \varepsilon_t,\ \varepsilon_t \sim iid(0, \Sigma)$$

$$\implies \quad y_t = B(L)\varepsilon_t = \varepsilon_t + B_1\varepsilon_{t-1} + B_2\varepsilon_{t-2} + ...$$

$$= \varepsilon_t + \Phi\varepsilon_{t-1} + \Phi^2\varepsilon_{t-2} + ...$$

But we need orthogonal shocks. Why?

So write $\varepsilon_t = Pv_t,\ v_t \sim iid(0, I)$, where P is Cholesky factor of Σ

$$\implies \quad y_t = Pv_t + (\Phi P)v_{t-1} + (\Phi^2 P)v_{t-2} + ...$$

ij'th IRF is the sequence of ij'th elements of $\{P,\ \Phi P\ \Phi^2 P, ...\}$ "Dynamic response of y_i to a one-σ shock in ε_j"