7 A Historical Perspective on Causal Inference in Macroeconometrics

The central objective in structural VAR analysis is to quantify causal relationships in the data. Before discussing the identification of causal relationships in structural VAR models, it is useful to review the precursors to structural VAR analysis. Our discussion traces how the focus of the literature has evolved from documenting lead-lag patterns in the data, as discussed in Sections 7.2–7.4, to quantifying unanticipated shifts in the data reflecting exogenous events, as discussed in Section 7.5. There are several approaches to constructing such exogenous shocks. We review the narrative approach to measuring exogenous policy shocks, the derivation of exogenous shocks from data-based counterfactuals, the construction of news shocks from macroeconomic announcements, and the measurement of shocks to financial market expectations. The definition of exogenous shocks was generalized with the introduction of the structural VAR framework, as discussed in Section 7.6. The latter approach is based on decomposing fluctuations in the data that cannot be predicted based on past data into mutually uncorrelated exogenous shocks with economic interpretation that need not be directly observable. As we trace the evolution of this literature, we also formally introduce the concepts of predeterminedness, strict exogeneity, and Granger causality, highlighting the extent to which each approach relies on these concepts.

7.1 A Motivating Example

The need for structural models in studying causal relationships between economic time series is best illustrated by the debate about causality from monetary aggregates to national income in the 1960s and 1970s. It had long been observed that money growth and income growth in the United States were positively correlated. Based on a careful review of the historical evidence, Friedman and Schwartz (1963) in their *Monetary History of the United States* concluded that changes in money growth are causing changes in income growth (an obvious implication being that the Federal Reserve should pursue

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

a constant money growth rule to stabilize the business cycle). This position evolved into a school of thought known as monetarism. Monetarism emphasizes the relation of the level of the money stock to the level of aggregate real economic activity (see Sims 1980b).

The monetarist position contrasted with the prevailing Keynesian wisdom that monetary policy was not nearly as important as fiscal policy in explaining economic fluctuations. Keynesians responded to Friedman and Schwartz by making the case that monetary aggregates were passive and that changes in the growth of monetary aggregates were endogenous responses to changes in real economic activity. Because higher output requires more "grease" in the form of money to keep the economy going, it is changes in real activity that cause endogenous changes in monetary aggregates through the money multiplier of the banking system (see Tobin 1970).

In its simplest form, monetarism is a statement about contemporaneous correlations between changes in real money stocks and real income. Because this correlation, while consistent with monetarist theory, is easy to explain away as a passive response of the money stock to changes in real activity, Friedman and Schwartz stressed the historical tendency for movements in the money stock (or its rate of change) to precede movements in aggregate activity. This additional implication of monetarist theory is harder to explain as a passive response of the money stock to changes in real activity and, hence, is considered a more challenging test of monetarist theory.

Amid this sometimes heated debate between monetarists and Keynesians, a new time series methodology emerged in the early 1970s that promised an answer to questions of causality and soon enjoyed considerable popularity. This methodology was developed by Clive Granger, among others, and the statistical tests in question became widely known as Granger causality tests (see Granger 1969; Sims 1972). More precisely, Granger's proposal was to test the null hypothesis of no (Granger) causality, with (Granger) causality being implied by the rejection of the null, as discussed in Chapter 2.

7.2 Granger Causality Tests for Covariance Stationary VAR Models

For expository purposes, consider the bivariate money-income autoregression

$$\begin{pmatrix} \Delta m_t \\ \Delta n_t \end{pmatrix} = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} + \begin{bmatrix} a_{11,1} & a_{12,1} \\ a_{21,1} & a_{22,1} \end{bmatrix} \begin{pmatrix} \Delta m_{t-1} \\ \Delta n_{t-1} \end{pmatrix} + \cdots + \begin{bmatrix} a_{11,p} & a_{12,p} \\ a_{21,p} & a_{22,p} \end{bmatrix} \begin{pmatrix} \Delta m_{t-p} \\ \Delta n_{t-p} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix},$$

where Δm_t denotes money growth and Δn_t denotes growth in national income. Then $\mathbb{H}_0 : a_{12,1} = \cdots = a_{12,p} = 0$ means that Δn_t does not Granger cause Δm_t, whereas $\mathbb{H}_0 : a_{21,1} = \cdots = a_{21,p} = 0$ means that Δm_t does not Granger

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

cause Δn_t. A Granger causality test can be conducted as a Wald test of the null hypothesis of no Granger causality. The possible results are:

1. Granger noncausality cannot be rejected in either direction.
2. Unidirectional Granger causality.
3. Bidirectional Granger causality.

Evidence of unidirectional Granger causality from money growth to income growth by many economists at the time would have been taken as evidence of Friedman and Schwartz being right about a causal role for money growth. Upon closer inspection, it is clear that Granger causality actually means that, on average over the sample period, including lags of one variable helps reduce the squared error in predicting another variable. It does not imply causality, only precedence. In other words, movements in one variable on average predate (or lead) movements in the other. The key difference between predictability and causality is that statistical precedence alone is not exploitable by policymakers. Separating cause and effect instead requires a structural model. This point is best illustrated by examples.

Example 1. A recent study by Shimmura and Yoshimura (2013) shows that roosters have internal clocks that make them crow before the sun comes up, even when kept in the dark. Hence, the crowing of a rooster helps predict the sunrise. If this were also a causal relationship, then a deliberate intervention such as strangling the rooster should prevent the sun from coming up. We know this not to be the case, so the predictive relationship is not causal.

In this example the data would have suggested Granger causality (precedence), yet it is clear that there is no causality in the sense we usually have in mind. The example was chosen to make the point that there is an obvious difference between precedence and causality. No one would confuse the two concepts in this example. Now consider a similar example drawn from economics.

Example 2. Most oil price increases have been followed by U.S. recessions. Hence, oil price increases cause recessions.

The difference is that now it is not immediately obvious whether the conclusion that oil price increases cause recessions is true or false, but it is clear that the logic that this statement appeals to in deriving its conclusion is flawed. The logical flaw is known as the *post hoc, ergo propter hoc* fallacy. In other words, precedence does not necessarily imply causality.

Granger and Newbold (1977, p. 225) acknowledge this drawback in the concept of Granger causality, but dismiss it: "Possibly cause is too strong a term, or one too emotionally laden, to be used. A better term might be temporally related, but since cause is such a simple term we shall continue to use it". Granger (1980, p. 333) elaborates that, "provided I define what I personally

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

mean by causation, I can use the term. I could, if I so wish, replace the word 'cause' throughout my lecture by some other words, such as 'oshkosh' or 'snerd', but what would be gained? It is like saying that whenever I use x, you would prefer me to use z."

This cavalier attitude has not gone unchallenged. For example, Leamer (1985, p. 284) in a discussion of Granger's work strongly objects to Granger taking control of the English language for his own purposes: "If I were to continue in that tradition I would propose that we henceforth refer to this notion of precedence by the wordpair: 'fool's causation'. This substitutes a loaded word 'fool' for the neutral 'Granger', just as causation has replaced the neutral precedence. Moreover, 'fool' is decidedly simpler than 'Granger' – it contains only four letters, one of which is repeated – and like 'cause', it is rather difficult to define precisely. One man's fool is another man's genius. My definition of a 'fool' would be a friend of mine living in San Diego." The last sentence is a reference to Granger who at the time was a professor at UC San Diego.

7.3 Granger Causality, Predeterminedness, and Exogeneity

The deeper reason why the distinction between predictability and causality matters is that predictive relationships need not be exploitable for policy purposes. It is useful to define more formally what we mean by a causal relationship and how causality relates to the concepts of Granger causality, predeterminedness, and strict exogeneity.[1]

7.3.1 Basic Concepts

The following discussion builds on Cooley and Leroy (1985). As before, let Δm_t denote money growth and Δn_t growth in national income. For expository purposes, consider the structural model:

$$\Delta m_t = \theta \Delta n_t + \beta_{11} \Delta m_{t-1} + \beta_{12} \Delta n_{t-1} + w_{1t},$$
$$\Delta n_t = \gamma \Delta m_t + \beta_{21} \Delta m_{t-1} + \beta_{22} \Delta n_{t-1} + w_{2t},$$

where w_{1t} and w_{2t} refer to mutually uncorrelated white noise innovations. Given this structural model, we can define the notions of predeterminedness and strict exogeneity.

Definition 1. *Δm_t is **predetermined** for Δn_t if $\theta = 0$.*

[1] Engle, Hendry, and Richard (1983) generalized the notion of exogeneity to possibly nonlinear models. In doing so they changed the terminology. Their notion of *weak exogeneity* in linear models corresponds to predeterminedness. Their notion of *strong exogeneity* in linear models reduces to the concept of strict exogeneity. Given our focus on linear models, we use the traditional definitions.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

Definition 2. *Δm_t is **(strictly) exogenous** with respect to Δn_t if $\theta = \beta_{12} = 0$ (so there is neither contemporaneous nor lagged feedback from Δn_t to Δm_t).*

The notion of strict exogeneity includes that of predeterminedness by construction. It is immediately apparent that the notion of predeterminedness captures what economists have in mind when referring to causality because a policymaker controlling Δm_t also controls Δn_t. Predeterminedness allows us to interpret a correlation between Δm_t and Δn_t as evidence of a causal effect of Δm_t on Δn_t, provided Δm_t is not correlated with any other predetermined variable that is not included in the bivariate model above. For example, if we think of the change in the price of oil as predetermined with respect to income growth and if this variable is correlated with Δm_t, the correlation between Δm_t and Δn_t can no longer be used to quantify the causal effect of Δm_t on Δn_t.

The notion of Granger causality may be defined based on the VAR(1) reduced-form representation of the structural model above:

$$\begin{aligned}\Delta m_t &= a_{11,1}\Delta m_{t-1} + a_{12,1}\Delta n_{t-1} + u_{1t},\\ \Delta n_t &= a_{21,1}\Delta m_{t-1} + a_{22,1}\Delta n_{t-1} + u_{2t}.\end{aligned}$$

Definition 3. *Δn_t fails to **Granger cause** Δm_t if $a_{12,1} = 0$.*

The first question is what Granger noncausality tells us about predeterminedness and strict exogeneity.

Predeterminedness. Granger noncausality is neither necessary nor sufficient for predeterminedness because

$$a_{12,1} \equiv \frac{\theta\beta_{22} + \beta_{12}}{1 - \theta\gamma} = 0 \quad \not\Rightarrow \quad \theta = 0.$$

Clearly, $\theta\beta_{22} = -\beta_{12}$ can hold for $\theta \neq 0$. Moreover,

$$\theta = 0 \quad \not\Rightarrow \quad a_{12,1} \equiv \frac{\theta\beta_{22} + \beta_{12}}{1 - \theta\gamma} = 0.$$

For example, $a_{12,1} \neq 0$ for $\theta = 0$ and $\beta_{12} \neq 0$. We conclude that one variable being Granger noncausal for the other does not imply that the latter variable is predetermined with respect to the former. Conversely, one variable being predetermined with respect to the other does not necessarily imply Granger noncausality from the latter variable to the former variable.

Next we turn to the relationship between strict exogeneity and Granger causality.

Strict exogeneity. Similarly,

$$a_{12,1} \equiv \frac{\theta\beta_{22} + \beta_{12}}{1 - \theta\gamma} = 0 \quad \not\Rightarrow \quad \theta = \beta_{12} = 0$$

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

because $\theta\beta_{22} = -\beta_{12}$ can hold for $\theta \neq 0$ and $\beta_{12} \neq 0$. Intuitively, Granger noncausality cannot imply exogeneity, because predeterminedness is part of the definition of exogeneity. However, strict exogeneity implies Granger noncausality because

$$\theta = \beta_{12} = 0 \quad \Rightarrow \quad a_{12,1} \equiv \frac{\theta\beta_{22} + \beta_{12}}{1 - \theta\gamma} = 0.$$

This means that by testing Granger noncausality we can test one of the implications of exogeneity. If we reject Granger noncausality, we also reject strict exogeneity. If we do not reject Granger noncausality, we learn nothing about whether strict exogeneity holds.

Sometimes the objective of conducting a Granger causality test is to establish that a variable is exogenous. Unfortunately, this result is precisely what a Granger noncausality test is unable to establish. Moreover, establishing exogeneity is useful if we want to justify exclusion restrictions in a structural model, but is stronger than what is needed to study the effects of policy interventions. For the latter purpose, predeterminedness suffices, which, however, is inherently untestable within the VAR framework.[2]

In practical terms, we conclude that even if money unidirectionally Granger causes income in a VAR model, this does not necessarily mean that the economy will grow faster, if the money supply is increased. This fact has not prevented many researchers to this day from misinterpreting Granger causality tests as genuine tests of causality. While they have learned not to refer to "causality" explicitly, and often acknowledge at some point that Granger causality does not imply causality, when it comes to the substantive results of their work, they simply replace the statement "X causes Y" by substantively identical statements such as "X is responsible for Y," "X explains Y," "X influences Y," "X is the source of Y," or "X has an effect on Y." This practice is incorrect and misleading, but it illustrates the lure of answering the causality question.

7.3.2 Granger Causality and Forward-Looking Behavior

Questions of causality are just as interesting in the context of financial economics. An important result is that financial asset prices will tend to Granger cause macroeconomic aggregates (but not the other way around) even in the absence of a causal relationship. The following example from Hamilton (1994, Example 11.1, pp. 306–307) illustrates this point. Recall that an investor who buys one share of a stock for the price P_t at date t, and on date $t + 1$ receives a

[2] Kilian and Vega (2011) provide an example of how high-frequency data on U.S. macroeconomic news may be used to address the question of the predeterminedness of energy prices with respect to U.S. macroeconomic aggregates.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

dividend D_{t+1} and sells the stock for P_{t+1}, receives an ex post rate of return of

$$r_{t+1} \equiv \frac{P_{t+1} + D_{t+1}}{P_t} - 1.$$

Hamilton postulates that the expected rate of return is a constant r such that

$$(1 + r)P_t = \mathbb{E}_t(P_{t+1} + D_{t+1}).$$

This first-order difference equation in P_t implies (after some algebra):

$$P_t = \mathbb{E}_t \left(\sum_{j=1}^{\infty} \left(\frac{1}{1+r} \right)^j D_{t+j} \right)$$

provided we rule out speculative bubbles in P_t. This means that the stock price today reflects the market's anticipation of future dividends. In other words, expectations of dividends drive the stock price in this model.

Next, Hamilton shows that P_t and D_t may be modeled as a bivariate VAR process. For concreteness, suppose that dividends follow the process

$$D_t = d + u_t + \delta u_{t-1} + v_t,$$

where d denotes a constant and u_t and v_t are mutually independent white noise processes. Hamilton observes that

$$\mathbb{E}_t(D_{t+j}) = \begin{cases} d + \delta u_t, & j = 1 \\ d, & j = 2, 3, \ldots \end{cases}$$

He substitutes this expectation of future dividends into the expression for P_t above:

$$\begin{aligned} P_t &= \mathbb{E}_t \left(\sum_{j=1}^{\infty} \left(\frac{1}{1+r} \right)^j d \right) + \mathbb{E}_t \left(\frac{1}{1+r} \delta u_t \right) \\ &= \frac{d}{r} + \frac{\delta u_t}{1+r} \quad \forall t, \end{aligned}$$

where $\sum_{j=1}^{\infty} \left(\frac{1}{1+r} \right)^j = \sum_{j=0}^{\infty} \left(\frac{1}{1+r} \right)^j - 1 = 1/r$. From

$$P_t = \frac{d}{r} + \frac{\delta u_t}{1+r}$$

it follows that

$$P_{t-1} = \frac{d}{r} + \frac{\delta u_{t-1}}{1+r} \quad \Leftrightarrow \quad \delta u_{t-1} = (1+r)P_{t-1} - (1+r)\frac{d}{r}.$$

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

Finally, Hamilton substitutes for δu_{t-1} in the expression for D_t:

$$\begin{aligned} D_t &= d + u_t + \delta u_{t-1} + v_t \\ &= d + u_t + (1+r)P_{t-1} - (1+r)\frac{d}{r} + v_t. \end{aligned}$$

Hence, the expressions for P_t and D_t can be written in VAR(1) format as:

$$\begin{pmatrix} P_t \\ D_t \end{pmatrix} = \begin{bmatrix} d/r \\ -d/r \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1+r & 0 \end{bmatrix} \begin{pmatrix} P_{t-1} \\ D_{t-1} \end{pmatrix} + \begin{pmatrix} \delta u_t/(1+r) \\ u_t + v_t \end{pmatrix}.$$

The VAR representation shows that D_t fails to Granger cause P_t, but P_t Granger causes D_t. This direction of Granger causality may seem surprising given that stock prices in the model depend on the expected path of dividends. Certainly, higher stock prices do not actually cause dividends to increase. Rather, this pattern of Granger causality arises because stock prices are determined by forward-looking agents who anticipate future dividends. This example provides some intuition for the tendency of time series that reflect forward-looking behavior such as asset prices or interest rates to be excellent predictors of macroeconomic time series such as real GDP or inflation. This clearly does not mean that these time series cause GDP or inflation to move. Rather, asset prices incorporate the market's assessment of where real GDP and inflation are headed.

7.3.3 Strict Exogeneity in Modern Macroeconomic Models

As discussed earlier, one of the few practical uses of Granger causality tests is for rejecting the strict exogeneity of economic variables. With the demise of traditional DSEMs and the emergence of SVAR models with fully endogenous variables, this question lost much of its urgency. The profession quickly accepted the notion that macroeconomic aggregates are unlikely to be strictly exogenous. One exception is macroeconomic models for small open economies. It has remained common to treat macroeconomic aggregates abroad as exogenously determined from the point of view of small open economies (see, e.g., Cushman and Zha 1997).

For a long time, the most promising candidate for an exogenous variable in macroeconomic models had been the price of crude oil. For example, Hamilton (2003) made the case that fluctuations in the price of crude oil are exogenous with respect to the U.S. economy to the extent that they are driven by politically motivated disruptions of oil production in the Middle East. This interpretation was subsequently questioned by Barsky and Kilian (2001) who challenged the exogeneity of some of the political events considered by Hamilton, and highlighted instabilities in the relationship between arguably exogenous oil supply disruptions and oil price increases. Subsequently, Kilian (2008a; 2008b) established that exogenous oil supply shocks in the Middle

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

East lack predictive power for the price of oil. Finally, Alquist, Kilian, and Vigfusson (2013) demonstrated that after 1973 the nominal price of oil is Granger caused by both U.S. and global macroeconomic aggregates, while the real price of oil is Granger caused by measures of global macroeconomic aggregates. At this point, there is no doubt that even the price of oil is not strictly exogenous, and a literature has evolved on modeling the endogeneity of oil prices both empirically and theoretically.

While strictly exogenous variables are hard to find in macroeconomics, the strict exogeneity assumption has survived in DSGE models which postulate that variables such as technology growth, money supply growth, or government spending growth are strictly exogenous. These variables are treated as exogenous mainly because of our inability or unwillingness to model them, not because there is evidence supporting the strict exogeneity assumption. Indeed, the endogeneity of U.S. money growth with respect to the real economy had been stressed by Keynesians already in the 1960s, and the endogeneity of government spending with respect to the business cycle is self-evident. Even the exogeneity of technology shocks is not obvious. For example, the theory of economic growth is devoted to making technology shocks endogenous with respect to the economy.

7.4 The Demise of Granger Causality Tests in Macroeconomics

As the profession developed a better understanding of the meaning of Granger causality tests and their statistical properties, the debate over whether money Granger causes income evolved. Initially, economists thought that bivariate Granger causality tests established that money leads income, but that result weakened once more variables were included in the VAR model. It also proved highly sensitive to different forms of detrending and to changes in the model specification. Moreover, it was shown that apparent Granger noncausality from money to income may simply reflect the omission of a third variable, whereas a finding of bivariate Granger causality may likewise reflect the omission of a third variable, calling into question even further the usefulness of Granger causality tests (see Lütkepohl 1982b).

By the late 1980s, the consensus was that money probably Granger causes income, but that this result does not prove monetarists right. At best it can be considered an empirical regularity that a macroeconomic model should be able to explain, not unlike a correlation pattern. As a result, the profession lost interest in Granger causality tests. Granger causality tests were replaced by the new idea that we are not interested in the contemporaneous correlation between money and income growth or the lead-lag pattern, but in the question of how income responds to unanticipated changes in money growth (also known as innovations or shocks). The hope was that these shocks would be exogenous, even if the underlying money growth time series was not.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

7.5 Responses to Unanticipated Changes in Money Growth

This new idea of focusing on unanticipated changes in money growth as measures of exogenous shocks emerged in four different guises: (1) an effort to identify exogenous changes in monetary policy based on the narrative approach to monetary policy; (2) efforts to measure the news (or surprise) component of monetary aggregates; (3) efforts to recover shocks to expectations from the prices of futures contracts; and (4) efforts to identify monetary policy shocks within the context of a structural VAR model.

7.5.1 The Narrative Approach

The narrative approach to identifying monetary policy shocks dates back to Friedman and Schwartz (1963). Not content to rely on statistical evidence only, Friedman and Schwartz used historical records to provide evidence for the existence of major swings in the real money stock which not only precede major swings in real activity but cannot themselves be explained as endogenous responses to changes in real activity. This approach was subsequently extended and formalized by Romer and Romer (1989), who introduced a dummy variable that takes on a value of 1 during periods when, according to the Romers' reading of the U.S. monetary policy records, the Federal Reserve exogenously tightened monetary policy, and zero otherwise. Their analysis involved a careful reading of statements of policymakers to determine whether a given change in monetary policy (say, an increase in the Federal Funds rate) represented an endogenous response to macroeconomic developments or an exogenous tightening of monetary policy.[3]

The original Romer dates included October 1947, September 1955, December 1968, April 1974, August 1978, and October 1979. Subsequently, they added December 1988 to this list. In related work, Kashyap, Stein, and Wilcox (1993) suggested February 1966 as another candidate and Oliner and Rudebusch (1995) proposed August 1988. To the extent that these Romer dummies, as they have come to be known, represent exogenous shocks, we may use distributed-lag models of the form

$$\Delta n_t = \nu + \sum_{i=0}^{m} \phi_i d_{t-i} + \varepsilon_t,$$

[3] The Romer shock measures are distinct from measures of the monetary policy stance that measure whether the Federal Reserve is leaning toward a contractionary or an expansionary policy. Examples of the latter measures include Boschen and Mills' (1995) quantitative dummy measure of the intensity of monetary policy stance (expressed on a scale of $\{-2, -1, 0, +1, +2\}$) and Bernanke and Mihov's (1998b) VAR-based index which is measured on a continuous scale. These measures of stance are not measures of exogenous shocks, but of policy actions. In other words, they are endogenous with respect to the state of the macroeconomy.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

where d_t denotes the dummy variable and Δn_t denotes income growth, to estimate the impulse responses directly by LS, where ε_t may be serially correlated and/or heteroskedastic, necessitating the use of robust standard errors in constructing t or Wald test statistics. The impulse response in the distributed-lag model simply is:

$$\frac{\partial \Delta n_t}{\partial d_{t-i}} = \frac{\partial \Delta n_{t+i}}{\partial d_t} = \phi_i$$

and confidence bands for each horizon i can be constructed based on each coefficient's robust standard errors. The number of lags determines the maximum horizon of the response functions. Note that the distributed-lag model may be viewed as a special case of the final-form representation of a dynamic simultaneous equations model (e.g., Lütkepohl 2005, section 10.2.2). For these impulse response estimates to capture the causal effects of policy interventions it must be the case that the dummies are truly exogenous, of course, and that they are uncorrelated with alternative sources of exogenous variation in Δn_t.

While the distributed-lag model is the simplest approach to estimating the responses in question, it is not the only one. In fact, Romer and Romer (1989), rather than estimating a distributed-lag model, report results based on the LS estimates of a single-equation model that includes additional lags of the dependent variable:

$$\Delta n_t = \nu + \sum_{i=0}^{m} \gamma_i d_{t-i} + \sum_{i=1}^{p} \beta_i \Delta n_{t-i} + \epsilon_t,$$

where ϵ_t may be heteroskedastic, but p has been chosen to render the error term serially uncorrelated. This alternative model can be motivated based on the premise that Δn_t and d_t are jointly determined by a structural VAR model of the form

$$\begin{aligned} d_t &= \theta \Delta n_t + \beta_{11} d_{t-1} + \beta_{12} \Delta n_{t-1} + w_{1t}, \\ \Delta n_t &= \gamma d_t + \beta_{21} d_{t-1} + \beta_{22} \Delta n_{t-1} + w_{2t}, \end{aligned}$$

where we have set $p = m = 1$ for expository purposes and suppressed all deterministic regressors. Romer and Romer (1989) postulate that (1) d_t is strictly exogenous with respect to Δn_t such that $\theta = 0$ and $\beta_{12} = 0$, and that (2) d_t is serially uncorrelated such that $\beta_{11} = 0$. Hence,

$$\begin{aligned} d_t &= w_{1t}, \\ \Delta n_t &= \gamma d_t + \beta_{21} d_{t-1} + \beta_{22} \Delta n_{t-1} + w_{2t}. \end{aligned}$$

One can compute $\partial \Delta n_{t+i} / \partial d_t$ from this restricted VAR model using the standard tools for impulse response analysis discussed in Chapter 4. This derivation

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

is not quite correct, however, in that d_t is a binary variable, which is at odds with the definition of the linear VAR model in Chapter 2.

One problem with the Romer dummies is that they are few in numbers, so the accuracy of the response estimates tends to be low. A second problem is that they make no allowance for the magnitude of the exogenous policy intervention. Third, they ignore episodes in which monetary policy expanded exogenously. The fourth and most important problem is that estimates of probit and logit models suggest that d_t is predictable based on past and/or expected macroeconomic aggregates and hence not exogenous with respect to the U.S. economy (see Shapiro 1994; Leeper 1997). This evidence suggests that Romer and Romer did not succeed in isolating the exogenous component of the change in monetary policy.

In response to this concern, Romer and Romer (2004) presented a refined measure of monetary policy shocks, which is constructed as the residual of a regression of the change in the intended federal funds rate around the dates of Federal Open Market Committee (FOMC) meetings on a set of predictors intended to purge the endogenous component of this series. If there are no meetings in a given month, the monetary policy shock is set to zero. The set of predictors used to distinguish exogenous shifts in monetary policy from policy responses to the expected state of the economy, includes the Federal Reserve's internal forecasts of inflation, real output growth, and the unemployment rate, as published in the Federal Reserve's Green Book. The resulting modified monetary policy shocks are available at monthly frequency for 1969m1-1996m12 and may take on positive, zero, or negative values. Constructing such modified monetary policy shocks, of course, creates a generated regressor problem that is typically ignored in applied work (see Pagan 1984). Quarterly measures of these monetary policy shocks are constructed by summing the monthly policy shocks by quarter.

The narrative approach has also been used for quantifying fiscal policy shocks. As far as government spending is concerned, Ramey and Shapiro (1998) introduce binary dummies for shocks to U.S. government spending associated with exogenous military buildups. Ramey (2011) refines this approach. Her work uses newspaper sources to quantify the changes in the expected present value of government spending associated with military buildups, resulting in a quantitative dummy variable that takes on nonzero values (measured on a continuous scale) on dates selected based on extraneous information and is zero otherwise.

On the government revenue side, Romer and Romer (2010) use the legislative records to construct dummies for tax changes. Their analysis allows them to separate legislated changes into those taken for reasons related to prospective economic conditions and those taken for exogenous reasons. Exogenous nominal tax changes are expressed as a fraction of nominal GDP in the quarter the change occurred, resulting in a quantitative dummy series.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

Fiscal dummies can be used for econometric analysis much the same way as monetary dummies. It is also common to include quantitative fiscal dummies as exogenous variables in VAR models (e.g., Ramey 2011; Mertens and Ravn 2012, 2013), as discussed in Chapter 15.

This literature has evolved further in recent years. A number of authors have defined fiscal policy shocks as an event consisting of multiple components, some of which are implemented immediately and some of which are left to be implemented in the future. An example is Alesina, Favero, and Giavazzi (2015). These components must be viewed in conjunction when assessing the response of the variable of interest. In other words, the fiscal policy shock is multidimensional rather than a scalar. Responses to such fiscal shocks may be constructed as the average difference between the simulated path of the variable of interest with and without all components of the plan implemented (see also Chapter 18).

A third area in which the narrative approach has been popular is the literature on oil supply shocks. Dotsey and Reid (1992) and Hoover and Perez (1994) introduce binary dummies for the dates of exogenous shortfalls in the production of oil in member countries of the Organization of Petroleum Exporting Countries (OPEC). Hamilton (2003) proposes quantitative dummies for measuring exogenous shortfalls in OPEC oil production, defined to be negative on a continuous scale on dates of exogenous political events in the Middle East and zero otherwise.

7.5.2 Exogenous Shocks Derived from Data-Based Counterfactuals

Although the narrative approach to identifying shocks has gained some prominence in measuring monetary policy shocks, oil supply shocks, and, more recently, fiscal policy shocks, it is but one example of methodologies aimed at identifying exogenous variation in the data. For example, Kilian (2008b) designs a methodology for quantifying the extent of exogenous variation in OPEC oil production from explicit counterfactuals about how oil production would have evolved in the absence of exogenous political events such as wars or revolutions. Related work includes Bastianin and Manera (2017). This approach results in a time series of the evolution of the exogenous component of OPEC oil production rather than a quantitative dummy. No estimation is involved, so there is no generated regressor problem. Changes in this exogenous series are serially uncorrelated, allowing the estimation of impulse responses from distributed-lag models as well as restricted VAR models. Another prominent example is the measure of exogenous technology shocks proposed by Basu, Fernald, and Kimball (2006). Their analysis controls for variation in capacity utilization in an effort to isolate the exogenous component of aggregate productivity.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

7.5.3 News Shocks

A closely related idea in the literature is that we can identify exogenous shocks by comparing the announcements of macroeconomic data releases with measures of what the market expected prior to the release. One advantage of this approach is that the news component of the announcement is unanticipated by construction. Another advantage is that we can obtain measures of news shocks at daily frequency, which makes this approach very useful when working with daily asset returns, r_t, for example. The news shock for a variable z_t after the announcement of its latest data release, denoted by $z_t^{\text{announcement}}$, is defined as

$$z_t^{\text{news}} \equiv z_t^{\text{announcement}} - \mathbb{E}_{t-1}(z_t^{\text{announcement}})$$

and may be standardized for better comparability across different types of news.

Of course, news shocks are only as good as the underlying measures of ex ante expectations. Typically, we obtain those expectations from market surveys taken the day before the announcement rather than from econometric models. In some cases, futures prices have been used as measures of expectations. There are data releases for about 30 different U.S. macroeconomic time series, including the latest industrial production data, housing starts, retail sales, construction spending, manufacturing orders, consumer confidence, initial unemployment claims, and other leading indicators of real activity. There also are news releases for monetary aggregates such as the federal funds target rate, allowing us to quantify monetary policy surprises. Given that news shocks are constructed at very high frequency, we can think of macroeconomic news shocks as plausibly exogenous shocks to the variables in question.

Again, the distributed-lag model

$$r_t = \nu + \sum_{i=0}^{m} \phi_i z_{t-i}^{\text{news}} + \varepsilon_t,$$

may be applied to estimate the impulse response

$$\frac{\partial r_{t+i}}{\partial z_t^{news}} = \phi_i.$$

Often the presumption is that asset prices should be informationally efficient in that news should be absorbed within the same day. That means that we do not need to worry about including lags of z_t^{news}:

$$r_t = \nu + \phi_0 z_t^{\text{news}} + \varepsilon_t.$$

When running regressions on news shocks, we sometimes stack the observations only for those time periods for which we have a news shock observation. This makes this type of model different from standard distributed-lag models. The sample size thus corresponds to the number of news shocks rather

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

than T. The model lends itself to t and Wald tests of whether news shocks matter. This distributed-lag model again may be viewed as a special case of the final form representation of a dynamic simultaneous equations model.

A good example for the literature on news shocks is Andersen, Bollerslev, Diebold, and Vega (2003). This paper studies the real-time price discovery in foreign exchange markets. The question is whether daily foreign exchange returns respond to the news component of macroeconomic data releases. The answer is yes. This finding is in striking contrast to the usual random walk result that monthly foreign exchange returns cannot be predicted based on monthly macroeconomic aggregates.[4]

Often it is difficult to find data on agents' expectations for the variable of interest. This fact has prompted the development of a closely related approach exploiting data from professional forecasters. For example, Kilian and Hicks (2013) define exogenous shocks to real GDP based on the revisions to professional real GDP forecasts. Similarly, Ramey (2011) proposes a measure of government spending shocks based on the errors made by professional forecasters.

It should be noted that news shock regressions tend to be based on daily (or even intra-daily) data, making it difficult to retain power when extrapolating to horizons longer than 20 business days. Thus, this framework is rarely used for characterizing the monthly or quarterly shocks of interest to macroeconomists. It has been used, however, to provide evidence for the assumption that monthly U.S. oil prices and gasoline prices are predetermined with respect to the U.S. macroeconomy (see Kilian and Vega 2011).

7.5.4 Shocks to Financial Market Expectations

An alternative approach to measuring exogenous shocks to expectations about policy variables or about other variables of interest is to rely on financial markets. In the absence of a risk premium, under standard assumptions the price of a futures contract of maturity h, F_t^h, in equilibrium equals the conditional expectation of the spot price, S_{t+h}. In practice, there is evidence of risk premia in most futures markets that drives a wedge between the futures price and the conditional expectation of the spot price. This fact suggests adjusting the futures price by an estimate of the risk premium, $\widehat{RP}_t^h$, to construct a time series of financial market expectations,

$$\mathbb{E}_t\left[S_{t+h}\right] = F_t^h - \widehat{RP}_t^h.$$

[4] It should be added that news shocks as defined in this literature are distinct from the more recent use of this term in macroeconomics, originating with Beaudry and Portier (2006) (see Chapter 10). The latter literature is about shifts in expectations, not about news properly defined as in this section. Clearly, news need not be linearly related to shifts in expectations.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

A generalization of this approach is discussed in Baumeister and Kilian (2016c). By comparing the market expectations for horizon h with the corresponding realizations, monthly time series of shocks to the market expectations may be obtained from financial market data (see Baumeister and Kilian 2016a).

One problem is that shocks to interest rate expectations over the course of a month, for example, are not necessarily tied to monetary policy decisions. They may occur for many reasons. Piazzesi and Swanson (2008) suggest that in measuring monetary policy shocks we may simply focus on changes in F_t^h from the day preceding a policy announcement to the following day. Under the premise that the risk premium evolves only slowly over time, changes in the risk premium from one day to the next are likely to be negligible. This allows the construction of a daily time series of monetary shocks without having to estimate the risk premium at daily frequency. For related discussion, see also Rudebusch (1998), Kuttner (2001), and Cochrane and Piazzesi (2002), among others. This approach, which may be extended to intra-daily frequencies, of course raises the question of how to recover the monthly and quarterly policy shocks of interest to macroeconomists. In practice, various proposals have been made to scale and aggregate high-frequency policy shocks for use in VAR models. These proposals also form the basis of the nonstandard VAR models of monetary policy discussed in Chapter 15.

7.5.5 Summary

This discussion shows that there are many creative ways of constructing measures of exogenous shocks that can be used to quantify causal relationships in the data. While most studies in this literature rely on least-squares estimates of the effects of these shocks from distributed-lag models or VARX models, some studies have used exogenous shocks as instruments. For example, Hamilton (2003) and Kilian (2008a, 2008b) report results for single-equation instrumental variable (IV) estimates based on exogenous OPEC oil supply shocks. The latter two studies draw attention to the problem that exogenous shocks may be weak instruments, invalidating conventional IV analysis. Similar issues may arise when including exogenous shock series in VAR models. A formal analysis of estimating VAR models with weak external instruments can be found in Montiel Olea, Stock, and Watson (2015a). This approach is discussed in Chapter 15.

7.6 Structural VAR Shocks

Although direct measures of policy shocks, news shocks based on macroeconomic announcements, shock measures based on financial market expectations, and estimates of shocks based on data-based counterfactuals have been

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

used extensively in applied work, the range of questions that can be analyzed with these tools is limited. By far the most common approach to estimating economically meaningful shocks has been to rely on structural vector autoregressions, building on the work of Sims (1980a).[5] The premise of this approach is that the DGP can be approximated by a structural VAR model of the form

$$B_0 y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + w_t, \tag{7.6.1}$$

where the deterministic terms have been suppressed, y_t is a $K \times 1$ vector of model variables, the model coefficients B_i, $i = 0, \ldots, p$, are $K \times K$ matrices, and the elements of the $K \times 1$ vector w_t are mutually uncorrelated white noise with nonsingular diagonal covariance matrix Σ_w, with one or more of the elements of w_t having a distinct economic interpretation. This model is recognizable as a special case of the DSEM model of Chapter 6, in which the coefficients of the lagged variables have been left unrestricted.

By pre-multiplying both sides of equation (7.6.1) by B_0^{-1}, we obtain the corresponding reduced-form VAR representation,

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t, \tag{7.6.2}$$

where $A_i = B_0^{-1} B_i$, $i = 1, \ldots, p$, and $u_t = B_0^{-1} w_t$, the estimation of which has been the subject of Chapters 2, 3, and 5.

7.6.1 The Identification Problem

It is readily apparent that knowledge of B_0^{-1} suffices to recover the parameters and shocks of the structural model (7.6.1), given a consistent estimate of the reduced-form model (7.6.2). The estimation of B_0^{-1} requires the user to impose additional identifying restrictions on B_0 or B_0^{-1} that can be motivated based on economic theory, institutional knowledge, or other external constraints on the structural model. Imposing these additional identifying restrictions allows one to decompose the reduced-form errors u_t into mutually uncorrelated structural shocks, w_t, with an economic interpretation.

It is important to keep in mind that without a clear economic interpretation of the elements of w_t, model (7.6.1) would not be structural. In particular, it is not enough for the elements of w_t to be mutually uncorrelated. Some early VAR applications overlooked this requirement and relied on ad hoc assumptions for identification that made no economic sense. Such atheoretical VAR models attracted strong criticism, spurring the development of more explicitly structural VAR models starting in the mid-1980s (see, e.g., Cooley and Leroy 1985).

[5] Although structural VAR models were not used in applied work prior to the 1980s, their antecedents can be traced back to the pioneering work of Mann and Wald (1943).

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

In response to ongoing questions about the validity of commonly used identifying assumptions, the structural VAR model literature has continuously evolved since the 1980s. The next chapters trace the evolution of this literature. We focus on alternative approaches to the identification of structural shocks within the framework of a reduced-form VAR model, highlighting the conditions under which each approach is valid and discussing potential limitations of commonly employed methods. For example, Chapter 8 focuses on identification by short-run restrictions. Chapter 10 reviews identification by long-run restrictions. Identification by sign restrictions is discussed in Chapter 13. Chapters 14 and 15 summarize alternative approaches of achieving identification by exploiting heteroskedasticity in the data, by using external instruments, or based on high-frequency changes in futures prices. For now we set these issues aside and focus on the interpretation of structural VAR shocks, assuming that an estimate of B_0^{-1} can be obtained.

7.6.2 The Relationship between Structural VAR Shocks and Direct Shock Measures

Knowledge of the mapping from the reduced-form VAR representation to the structural VAR representation allows us to quantify the structural shocks, given that $w_t = B_0 u_t$. Compared with direct measures of exogenous shocks of the type discussed in Section 7.5, a key difference is that the structural shocks in model (7.6.1) are in general unobservable and need not be associated with any one VAR model variable in particular, which greatly increases the range of applications of the VAR approach. This point is best illustrated by example. Consider a prototypical microeconomic model of the price and quantity of a commodity, in which the observables price (p_t) and quantity (q_t) are driven by latent demand shocks (w_t^d) and supply shocks (w_t^s). This model may be written as a structural VAR model of the form

$$\begin{bmatrix} b_{11,0} & b_{12,0} \\ b_{21,0} & b_{22,0} \end{bmatrix} \begin{pmatrix} q_t \\ p_t \end{pmatrix} = \begin{bmatrix} b_{11,1} & b_{12,1} \\ b_{21,1} & b_{22,1} \end{bmatrix} \begin{pmatrix} q_{t-1} \\ p_{t-1} \end{pmatrix} + \cdots + \begin{bmatrix} b_{11,p} & b_{12,p} \\ b_{21,p} & b_{22,p} \end{bmatrix} \begin{pmatrix} q_{t-p} \\ p_{t-p} \end{pmatrix} + \begin{pmatrix} w_t^s \\ w_t^d \end{pmatrix},$$

or more compactly as

$$B_0 u_t = w_t \quad \Longleftrightarrow \quad \begin{bmatrix} b_{11,0} & b_{12,0} \\ b_{21,0} & b_{22,0} \end{bmatrix} \begin{pmatrix} u_t^q \\ u_t^p \end{pmatrix} = \begin{pmatrix} w_t^s \\ w_t^d \end{pmatrix},$$

where u_t is the error of the reduced-form VAR representation, which also may be interpreted as the unanticipated change in p_t and q_t caused by the structural

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

shocks in period t. This expression may be rearranged as

$$u_t = B_0^{-1} w_t \quad \Longleftrightarrow \quad \begin{pmatrix} u_t^q \\ u_t^p \end{pmatrix} = \begin{bmatrix} b_0^{11} & b_0^{12} \\ b_0^{21} & b_0^{22} \end{bmatrix} \begin{pmatrix} w_t^s \\ w_t^d \end{pmatrix}.$$

In this model, the observables price and quantity are simultaneously determined by demand and supply shocks, allowing us to express the unpredictable component of the data, u_t^q and u_t^p, as a weighted average of the w_t^s and w_t^d with the weights provided by the rows of B_0^{-1}, as illustrated by the last equation.

By construction, each structural shock affects the price and the quantity of the commodity in question simultaneously. Only if the model is recursive such that $b_0^{12} = b_{12,0} = 0$, rendering the quantity variable predetermined with respect to the price, is there a direct link between one of the structural shocks and one of the variables, as assumed in Section 7.5.

7.6.3 Causality in Structural VAR Models

The ultimate question of interest in structural VAR analysis is not the evolution of the structural shocks, but their causal effects on the model variables, as captured by impulse response functions (see Chapter 4). The nature of this link is best illustrated in the context of the model of demand and supply.

Structural shocks within this framework may be interpreted as exogenous shifts of the demand and supply curves in the underlying economic model. For example, a positive demand shock could be represented as an exogenous shift of the demand curve to the right along the supply curve. Thus, a demand shock traces out the slope of the supply curve, as captured by $b_{12,0}$. One of the requirements for a causal interpretation of the responses of future values of price and quantity to a demand or supply shock, respectively, in the current period, is that these shocks are mutually uncorrelated. Obviously, if w_t^d were correlated with w_t^s, both curves would shift, making it impossible to interpret the resulting changes in prices and quantities as the causal effect of either a demand shock or a supply shock.

Treating the demand and supply shocks in this example as exogenous makes sense, because they cannot be predicted based on past data and are not correlated with one another. Implicitly, we are assuming that the observed price and quantity data are generated by these two structural shocks only. This assumption is reasonable if the structural VAR model is correctly specified. There are counterexamples, however. For example, if the price and quantity data underlying this model were determined by the actions of economic agents whose expectations of future prices and quantities differ from the predictions implied by this VAR model, this particular structural VAR model would be invalid. This situation is discussed in Chapter 17. In some cases, this problem may be overcome by the use of suitable external instruments that embody the missing

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008

information, as discussed in Chapter 15, or by specifying VAR models that allow for more information, as discussed in Chapters 16 and 17.

Structural VAR models tend to involve much richer structures than this stylized model of demand and supply, of course, but this example illustrates how structural VAR models may be used to quantify causal relationships in the data.

Downloaded from https://www.cambridge.org/core. Boston College, on 30 Jan 2020 at 18:45:19, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108164818.008