2 Macroeconomic Methodology

2.1 Macro Theoretical Models and the Role of Theory

2.1.1 Ingredients of Models

Broadly speaking, an economy consists of people making and carrying out decisions and interacting with each other through markets. Theories provide explanations of how the decisions are made and how the markets work. The ingredients of a theory include the choice of the decision-making units, the decision variables and objective function of each unit, the constraints facing each unit, and the amount of information each unit has at the time the decisions are made. Possible constraints include budget constraints, technological constraints, direct constraints on decision variables, and institutional or legal constraints. If expectations of future values affect current decisions, another ingredient of a theory is an explanation of how expectations are formed.

A theory of how markets work should explain who sets prices and how they are set. If there is the possibility of disequilibrium in certain markets, the theory should explain how quantities are determined each period and why it is that prices are not set to clear the markets. Institutional constraints may play an important role in some markets.

In macroeconomics there are also a number of adding-up constraints that should be met. In particular, balance-sheet and flow-of-funds constraints should be met. An asset of one person is a liability of someone else, and income of one person in a period is an expenditure of someone else in the period. These two constraints are not independent, since any deviation of income from expenditure for an individual in a period corresponds to a change in at least one of his or her assets or liabilities.

2.1.2 The Traditional Role of Theory

An important issue in the construction of a model is the role that one expects theory to play. If the aim is to use the theoretical model to guide the specification of an empirical model, the issue is how many restrictions one

can expect theory to provide regarding the specification of the equations to be estimated. In practice, the primary role of theory has been to choose the variables that appear with nonzero coefficients in each equation. (Stated another way, the primary role of theory has been to provide "exclusionary" restrictions on the model, that is, to provide a list of variables *not* to include in each equation.) In most cases theory also chooses the signs of the coefficients. Much less often is theory used to decide things like the functional forms of the estimated equations and the lengths of the lag distributions. (This is not to say that theory could not be used for such purposes, only that it generally has not been.) This role of theory—the choice of the variables to include in each equation—will be called the "traditional" role or approach.

An interesting question within the traditional approach is whether theory singles out one variable per equation as the obvious dependent or "left-hand-side" (LHS) variable, where the other variables are then explanatory or "right-hand-side" (RHS) variables. In this way of looking at the problem, the LHS variable is the decision variable and the RHS variables are the determinants of the decision variable. If the theoretical problem is to explain the decisions of agents, this way seems natural. Each equation is a derived decision equation (derived either in a maximization context or in some other way) with a natural LHS variable. The alternative way of looking at the problem is that theory treats all variables in each equation equally. These two interpretations have important implications for estimation. In particular, full information maximum likelihood (FIML) treats all variables equally, whereas two-stage least squares (2SLS) and three-stage least squares (3SLS) require an LHS variable to be chosen for each equation before estimation (see, for example, Chow 1964). One might thus be inclined to choose 3SLS over FIML under the first interpretation, although there are other issues to consider in this choice as well. This issue is discussed in more detail in Section 6.3.4, where FIML and 3SLS are compared. For the remainder of this chapter it will be assumed that within the traditional approach the LHS variable is also chosen.

2.1.3 The Hansen-Sargent Approach and Lucas's Point

An alternative role for theory is exemplified by the recent work of Hansen and Sargent (1980). In this work the aim is to estimate the parameters of the objective functions of the decision-making units. In the traditional approach these parameters are never estimated. The parameters of the derived decision equations (rules) are estimated instead, where these parameters are functions

of the parameters of the objective function and other things. The Hansen-Sargent approach imposes many more theoretical restrictions on the data than does the traditional approach, especially considering that the traditional approach imposes very few restrictions on the functional forms and the lag structures of the estimated decision equations.

The advantage of the Hansen-Sargent approach is that it estimates structural parameters rather than combinations of structural parameters and other things. The problem with estimating combinations is that if, say, one wants to examine the effects of changing an exogenous variable on the decision variables, there is always the possibility that this change will change something in the combinations. If so, then it is inappropriate to use the estimated decision equations, which are based on fixed estimates of the combinations, to examine the effects of the change. This is the point emphasized by Lucas (1976) in his classic article. (Note that the validity of the point does not depend on expectations being rational. Even if expectations are formed in rather naive ways, it may still be that the coefficients of the decision equations are combinations of things that change when an exogenous variable is changed.)

There are two disadvantages of the Hansen-Sargent approach, one that may be temporary and one that may be more serious. The temporary disadvantage is that it is extremely difficult to set up the problem in such a way that the parameters can be estimated, especially if there is more than one decision variable or if the objective function is not quadratic. Very restrictive assumptions have so far been needed to make the problem tractable. This disadvantage may gradually be lessened as more tools are developed. At the present time, however, this approach is a long way from the development of a complete model of the economy.

A potentially more serious disadvantage, at least as applied to macroeconomic data, is the possibility that the approach imposes restrictions on the data that are poor approximations. Macroeconomic data are highly aggregated, and it is obviously restrictive to assume that one objective function pertains to, say, the entire household sector or the entire firm sector. Although both the traditional approach and the Hansen-Sargent approach are forced to make assumptions like this when dealing with macroeconomic data, the Hansen-Sargent approach is much more restrictive. If because of aggregation problems the assumption that a sector behaves by maximizing an objective function is not correct, models based on both approaches will be misspecified. This misspecification may be more serious for models based on the Hansen-Sargent approach because it uses the assumption in a much stronger way. To

put it another way, by not requiring that a particular objective function be specified, the traditional approach may be more robust to errors regarding the maximization assumption.

It is difficult to argue against the Hansen-Sargent approach without sounding as if one is in favor of the use of ad hoc theory to explain macroeconomic data. Arguments against theoretical purity are generally not well received in the economics profession. There are, however, as just discussed, different degrees to which theory can be used to guide econometric specifications. There is a middle ground between a completely ad hoc approach and the Hansen-Sargent approach, namely what I have called the traditional approach. An example of this approach is given in Chapters 3 and 4.

It should also be noted that the Hansen-Sargent approach can be discussed without reference to how expectations are formed. It is typically assumed within this approach that expectations are rational, but this is not a necessary assumption. It is clearly possible within the context of a maximization problem to assume that expectations of the future variable values that are needed to solve the problem are formed in simple or naive ways. The possible problems with the Hansen-Sargent approach discussed earlier exist independently of the expectational assumptions that are used. The problems are perhaps potentially more serious when the rational expectation assumption is used because of the tighter theoretical restrictions that are implied, but this is only a matter of degree. The treatment of expectations is discussed in Section 2.2.2.

Whether the Hansen-Sargent approach will lead to better models of the economy is currently an open question. As noted in Chapter 1, a major theme of this book is that it should be possible in the long run to decide questions like this using methods like the one discussed in Chapter 8. The method in Chapter 8 allows one to compare different models in regard to how well they approximate the true structure. If the Hansen-Sargent approach leads eventually to the construction of complete models of the economy, it should be possible to compare these models to models based on the traditional approach.

If because of the limitations just discussed the Hansen-Sargent approach does not lead to econometric models that are good approximations, this does not invalidate Lucas's point (1976). The point is a logical one. If parameters that are taken to be constant change when an exogenous variable is changed, the estimated effects of the change are clearly in error. The key question for any given experiment with an econometric model is the likely size of this error. There are many potential sources of error, and even the best economet-

ric model in the future (as judged, say, by the method in Chapter 8) will be only an approximation to the structure. It may be that for many experiments the error from the Lucas point is quite small. The question is how much the parameters of estimated decision equations, such as consumption and labor supply equations of the household sector, change when a government policy variable changes. For many policy variables and equations these changes may not be very great. The errors in the multipliers that result from not accounting for the parameter changes may be much smaller than, say, the errors that result from aggregation. At any rate, how important the Lucas point is quantitatively is currently an open question.

One encouraging feature regarding the Lucas point is the following. Assume that for an equation or set of equations the parameters change considerably when a given policy variable changes. Assume also that the policy variable changes frequently. In this case the method in Chapter 8 is likely to weed out a model that includes this equation or set of equations. The model is obviously misspecified, and the method should be able to pick up this misspecification if there have been frequent changes in the policy variable. It is thus unlikely that a model that suffers from the Lucas criticism will be accepted as the best approximation of the structure.

One may, of course, still be misled regarding the Lucas point if the policy variable has changed not at all or very little in the past. In this case the model will still be misspecified, but the misspecification has not been given a chance to be picked up in the data. The model may thus be accepted when in fact it is seriously misspecified with respect to the effects of the policy variable on the endogenous variables. One should thus be wary of drawing conclusions about the effects of seldom-changed policy variables unless one has strong reasons for believing that the Lucas point is not quantitatively important for the particular policy variable in question.

2.1.4 The Sims Approach

Another role for theory in the construction of empirical models has been stressed recently by Sims (1980). This role is at the opposite end of the spectrum from that advocated by Hansen and Sargent—namely, it is very limited. Sims does not trust even the exclusionary restrictions imposed by the traditional approach; he argues instead for the specification of vector autoregressive equations, where each variable is specified to be a function of its own lagged values and the lagged values of other variables. (An important early study in this area is that of Phillips 1959.) Although this approach imposes

some restrictions on the data—in particular, the number of variables to use, the lengths of the lags, and (sometimes) cross-equation restrictions on the coefficients—the restrictions are in general less restrictive than the exclusionary ones used by the traditional approach.

Although it is again an open question whether Sims's approach will lead to better models, it should be possible to answer this question by comparing models based on this approach to models based on other approaches. Some results that bear on this question are presented in this book. The method in Chapter 8 is used to compare my US model to two vector autoregressive models. The vector autoregressive models are presented in Section 5.2, and the comparison is discussed in Section 8.5.

2.1.5 Long-Run Constraints

In much macroeconomic modeling in which theory is used, various long-run constraints are imposed on the model. Consider, for example, the question of the long-run trade-off between inflation and unemployment. Economists with such diverse views as Tobin and Lucas seem to agree with the Friedman-Phelps proposition that there is no long-run trade-off. (See Tobin 1980, p. 39, and Lucas 1981, p. 560. For the original discussion of the Friedman-Phelps proposition see Friedman 1968 and Phelps 1967.) Accepting this proposition clearly colors the way in which one thinks about macroeconomic issues. Lucas, for example, points out that much of the recent work in macroeconomic theory has been concerned with trying to reconcile this long-run proposition with the observed short-run fluctuations in the economy (1981, p. 561). The imposition of long-run constraints of this type clearly has important effects on the entire modeling exercise, including the modeling of the short run.

Although it is difficult to argue this in the abstract, my feeling is that long-run constraints may be playing too much of a role in recent macroeconomic work. Consider the two possible types of errors associated with a particular constraint. The first is that an incorrect constraint is imposed. This error will lead to a misspecified model, and the misspecification may be large if the constraint has had important effects on the specification of the model and if it is a poor approximation. The second type of error is that a correct constraint is not imposed. Depending on the setup, this type of error may not lead to a misspecified model, but only one in which the coefficient estimates are inefficient. At any rate, it is my feeling that the first type of error may be more serious in practice than the second type, and if this is so, long-run

constraints should be imposed with considerable caution. It is not obvious, for example, that the assumption of no long-run trade-off between inflation and unemployment warrants so much confidence that it should be imposed on models, given the severe restrictions that it implies.

This argument about long-run constraints will be made clearer in Section 3.1.6 in the discussion of my theoretical model. Again, however, this issue of the imposition of long-run constraints can be tested (in the long run) by comparing models based on different constraints.

2.1.6 Theoretical Simulation Models

With the growth of computer technology there has been an increase in the number of theoretical models that are analyzed by simulation techniques. The main advantage of using these techniques is that much larger and more complicated models can be specified; one need not be restricted by analytic tractability in the specification of the model. A disadvantage of using the techniques is that the properties of the model may depend on the particular set of parameters and functions chosen for the simulation, and one may get a distorted picture of the properties. Although one can guard against this situation somewhat by performing many experiments with different sets of parameters and functions, simulation results are not a perfect substitute for analytic results.

The relationship between simulation exercises and empirical work is not always clearly understood, and it will be useful to consider this issue. If simulation techniques are merely looked upon as a substitute for analytic techniques when the latter are not feasible to use, then the relationship between simulation exercises and empirical work is no different from the relationship between analytic exercises and empirical work. The results of analyzing theoretical models are used to guide empirical specifications, and it does not matter how the theoretical model is analyzed. An example of the use of simulation techniques in this way is presented in this book. The theoretical model discussed in Chapter 3 is analyzed by simulation techniques, and the results from this model are used to guide the specification of the econometric model in Chapter 4. Had it been feasible to analyze the model in Chapter 3 by analytic techniques, this would have been done, and provided no new insights about the model were gained from this, the econometric specifications in Chapter 4 would have been the same. In this way of looking at the issue, the difference between simulation and analytic techniques is not important: the methodology is really the same in both cases.

Note with respect to empirical work that the type of theoretical simulation model just discussed is not an end in itself; it is merely a stepping-stone to the specification of the equations to be estimated. The data are used in the estimation and analysis of the derived empirical model (derived in a loose sense—see Section 2.2), not in the theoretical model itself. This type of theoretical simulation model is quite different from the type that has come to be used in the field of applied general equilibrium analysis. A good discussion of the methodology of this field is contained in Mansur and Whalley (1981), and it will be useful to review this methodology briefly to make sure there is no confusion between it and the methodology generally followed in macroeconomics.

There are two main steps in the construction of an applied general equilibrium model. The first is to construct for a given period (usually a particular year) a "benchmark equilibrium data set," which is a collection of data in which equilibrium conditions of an assumed underlying equilibrium model are satisfied. Considerable data adjustment is needed in this step because the existing data are generally not detailed enough (and sometimes not conceptually right) for a general equilibrium model. The data, for example, may not be mutually consistent in the sense that the model equilibrium conditions are not satisfied in the data. Most benchmark equilibrium data sets satisfy the following four sets of equilibrium conditions: (1) demand equals supplies for all commodities, (2) nonpositive profits are made in all industries, (3) all domestic agents (including the government) have demands that satisfy their budget constraints, and (4) the economy is in zero external balance. Condition (3) usually involves treating the residual profit return to equity as a contractual cost.

The second step is to choose the functional forms and parameter values for the model. These are chosen in such a way that the model is "calibrated" to the benchmark equilibrium data set. The fundamental assumption involved in this calibration is that the economy is in equilibrium in the particular year. The restriction on the parameter values is that they replicate the "observed equilibrium" as an equilibrium solution of the model. The values are determined by solving the equations that represent the equilibrium conditions of the model, using the data on prices and quantities that characterize the benchmark equilibrium. Depending on the functional forms used, the observed equilibrium may not be sufficient to determine uniquely the parameter values. If the values are not uniquely determined, some of them must be chosen ahead of time (that is, before the model is solved to get the other values). The values chosen ahead of time are generally various elasticities of

substitution; they are often chosen by searching the literature for estimated values.

Once the parameters are chosen, the model is ready to be used for policy analysis. Various exogenous variables can be changed, and the model can be solved for these changes. The differences between the solution values and the values in the data set are the estimates of the effects of the policy change. These estimates are general equilibrium estimates in the sense that the entire general equilibrium model is solved to obtain them.

The difference between this second type of theoretical simulation model and the first type should be clear. The second type is an end in itself with respect to empirical work: models of this type are used to make empirical statements. The main problem with this methodology, as is well known by people in the field, is that there is no obvious way of testing whether the model is a good approximation to the truth. The models are not estimated in the usual sense, and there is no way to use a method like the one in Chapter 8 to compare alternative models. Each model fits the data set perfectly, usually with room to spare in the sense that many parameter values are typically chosen ahead of time. This is contrasted with models of the first type, which can be indirectly tested by testing the empirical models that are derived from them (see the discussion in Section 2.3).

It is unclear at this stage whether the applied general equilibrium models will become more like standard econometric models and thus more capable of being tested or whether they will remain in their current "quasi-empirical" state. Whatever the case, the main point for this book is that the methodology followed here is quite different from the methodology currently followed in applied general equilibrium analysis.

2.2 The Transition from Theoretical to Econometric Models

The transition from theoretical models to empirical models is probably the least satisfying aspect of macroeconomic work. One is usually severely constrained by the quantity and quality of the available data, and many restrictive assumptions are generally needed in the transition from the theory to the data. In other words, considerable "theorizing" occurs at this point, and it is usually theory that is much less appealing than that of the purely theoretical model. Many examples of this will be seen in Chapter 4 in the discussion of the transition from the theoretical model in Chapter 3 to the econometric model in Chapter 4. This section contains a general discussion of the steps that are usually followed in the construction of an econometric model.

2.2.1 Step 1: Data Collection and the Choice of Variables and Identities

The first step is to collect the raw data, create the variables of interest from the raw data, and separate the variables into exogenous variables, endogenous variables explained by identities, and endogenous variables explained by stochastic equations. The data should match as closely as possible the variables in the theoretical model. In macroeconomic work this match is usually not very close because of the highly aggregated nature of the macro data. Theoretical models are usually formulated in terms of individual agents (households, firms, and the like), whereas the macro data pertain to entire sectors (household, firm, and the like). There is little that can be done about this problem, and for some it calls into question the usefulness of using theoretical models of individual agents to guide the specification of macroeconometric models. It may be, in other words, that better macroeconometric models can be developed using less micro-based theories. This is an open question, and it is another example of an issue that can be tested in the long run by comparing different models.

There are many special features and limitations of almost any data base that one should be aware of, and one of the most important aspects of macroeconometric work, perhaps the most important, is to know one's data well. Knowledge of how to deal with data comes in part through experience and in part from reading about how others have done it; it is difficult to learn in the abstract. Appendixes A and B of this book provide an example of the collection of the data for my model.

It is important, if possible, to have the data meet the adding-up constraints that were mentioned at the beginning of this chapter. In addition to such obvious things as having the data satisfy income identities, it is useful to have the data satisfy balance-sheet constraints. For the US data, this requires linking the data from the Flow of Funds Accounts to those from the National Income and Product Accounts. This is discussed in Chapter 4 and in Appendix A. The linking of these two data bases is a somewhat tedious task and is a good example of the time-consuming work that is involved in the collection of data.

The data base may be missing observations on variables that are essential for the construction of the model. In such cases, rather than giving up, it may be possible to construct estimates of the missing data. If, for example, the data for a particular variable are annual, whereas quarterly data are needed, it may be possible, using related quarterly variables, to create quarterly data from the annual data by interpolating. There are also more sophisticated procedures

for constructing missing observations (see, for example, Chow and Lin 1971). Appendix B provides a number of examples of the construction of missing data for my multicountry model.

Although it is easiest to think of the division of endogenous variables into those determined by stochastic equations and those determined by identities as being done in the first step, the choice of identities is not independent of the choice of explanatory variables in the stochastic equations. If a given explanatory variable is not exogenous and is not determined by a stochastic equation, it must be determined by an identity. It is thus not possible to list all the identities until the stochastic equations are completely specified.

2.2.2 Step 2: Treatment of Unobserved Variables

Most theoretical models contain unobserved variables, and one of the most difficult aspects of the transition to econometric specifications is dealing with these variables. Much of what is referred to as the "ad hoc" nature of macroeconomic modeling occurs at this point. If a theoretical model is explicit about the determinants of the unobserved variables and if the determinants are observed, there is, of course, no real problem. The problem is that many models are not explicit about this, and so "extra" modeling or theorizing is needed at this point.

Expectations

The most common unobserved variables in macroeconomics are expectations. A common practice in empirical work is to assume that expected future values of a variable are a function of the current and past values of the variable. The current and past values of the variable are then used as "proxies" for the expected future values. Given the importance of expectations in most models, it will be useful to consider this procedure in some detail.

Consider first the following example:

$$(2.1) \qquad y_t = \alpha_0 + \alpha_1 E_{t-1} x_{t+1} + u_t,$$

where $E_{t-1}x_{t+1}$ is the expected value of x_{t+1} based on information through period $t-1$. A typical assumption is that $E_{t-1}x_{t+1}$ is a function of current and past values of x:

$$(2.2) \qquad E_{t-1}x_{t+1} = \lambda_1 x_t + \lambda_2 x_{t-1} + \ \ldots \ + \lambda_n x_{t-n+1},$$

where it is assumed that x_t is observed at the beginning of period t. Given (2.2), two procedures can be followed to obtain an estimatable equation. One is to substitute (2.2) into (2.1) and simply regress y_t on the current and past values of x. (Other variables can also be used in 2.2 and then substituted into 2.1. If, say, z_t affects $E_{t-1}x_{t+1}$, then z_t would be used as an explanatory variable in the y_t regression.) A priori restrictions on the λ_i coefficients (that is, on the shape of the lag distribution) are sometimes imposed before estimation. Lagged values of time series variables tend to be highly correlated, and it is usually difficult to get estimates of lag distributions that seem sensible without imposing some restrictions. If no restrictions are imposed on the λ_i coefficients, α_1 cannot be identified.

The other procedure is to assume that the lag distribution is geometrically declining, in particular that $\lambda_i = \lambda^i$, $i = 1, \ldots, \infty$. Given this assumption, one can derive the following equation to estimate:

$$(2.3) \qquad y_t = \alpha_0(1-\lambda) + \lambda\alpha_1 x_t + \lambda y_{t-1} + u_t - \lambda u_{t-1}.$$

The coefficient of the lagged dependent variable in this equation, λ, is the coefficient of the lag distribution. It appears both as the coefficient of the lagged dependent variable and as the coefficient of u_{t-1}, and although this restriction should be taken into account in estimation work, it seldom is. Sometimes equations like (2.3) are estimated under the assumption of serial correlation of the error term (that is, an assumption like $v_t = \rho v_{t-1} + \epsilon_t$, where v_t denotes the error term in 2.3), but this is not the correct way of accounting for the λ restriction.

There is a nonexpectational model that leads to an equation similar to (2.3), which is the following simple lagged adjustment model. Let y_t^* be the "desired" value of y_t, and assume that it is a linear function of x_t:

$$(2.4) \qquad y_t^* = \alpha_0 + \alpha_1 x_t.$$

Assume next that y_t only partially adjusts to y_t^* each period, with adjustment coefficient γ:

$$(2.5) \qquad y_t - y_{t-1} = \gamma(y_t^* - y_{t-1}) + u_t.$$

Equations (2.4) and (2.5) can be combined to yield

$$(2.6) \qquad y_t = \lambda\alpha_0 + \lambda\alpha_1 x_t + (1-\gamma)y_{t-1} + u_t.$$

Equation (2.6) is in the same form as (2.3) except for the restriction on the error term in (2.3). As noted earlier, the restriction on the error term in (2.3) is

usually ignored, which means that in practice there is little attempt to distinguish between the expectations model and the lagged adjustment model. It may be for most problems that the data are not capable of distinguishing between the two models. The problem of distinguishing between the two is particularly difficult if the u_t error terms in (2.1) and (2.5) are assumed to be serially correlated, because in this case the differences in the properties of the error terms in the derived equations (2.3) and (2.6) are fairly subtle. At any rate, it is usually the case that no attempt is made to distinguish between the expectations model and the lagged adjustment model.

Two other points about (2.3) should be noted. First, if there is another variable in the equation, say z_t, the implicit assumption that is being made when this equation is estimated is that the expectations of z are formed using the same coefficient λ that is used in forming the expectations of x. In other words, the shape of the two lag distributions is assumed to be the same. This may be, of course, a very restrictive assumption. Second, if there is another future expected value of x in (2.1), say $\alpha_2 E_{t-1}x_{t+2}$, and if this expectation is generated as

$$E_{t-1}x_{t+2} = \lambda E_{t-1}x_{t+1} + \lambda^2 x_t + \lambda^3 x_{t-1} + \ldots, \tag{2.7}$$

then (2.3) is unchanged except for a different interpretation of the coefficient of x_t. The coefficient in this case is $\lambda(\alpha_1 + 2\alpha_2\lambda)$ instead of $\lambda\alpha_1$. The same equation would be estimated in this case, although it is not possible to identify α_1 and α_2.

It should be clear that this treatment of expectations is somewhat unsatisfying. Agents may look at more than merely the current and past values of a variable in forming an expectation of it, and even if they do not, the shapes of the lag distributions may be quite different from the shapes usually imposed in econometric work. The treatment of expectations is clearly an important area for future work. An alternative treatment to the one just presented is the assumption that expectations are rational. This means that agents form expectations by first forming expectations of the exogenous variables (in some manner that must be specified) and then solving the model using these expectations. The predicted values of the endogenous variables from this solution are the expected values. The assumption of rational expectations poses a number of difficult computational problems when one is dealing with large-scale nonlinear models, but many of these problems are now capable of solution. Chapter 11 discusses the solution and estimation of rational expectations models.

It is by no means obvious that the assumption that expectations are rational

is a good approximation to the way that expectations are actually formed. The assumption implies that agents know the model, and this may not be realistic for many agents. It would be nice to test assumptions that are in between the simple assumption that expectations of a variable are a function of its current and past values and the assumption that expectations are rational. One possibility is to assume that expectations of a variable are a function not only of its current and past values but also of the current and past values of other variables. To implement this, the variable in question could be regressed on a set of variables and the predicted values from this regression taken to be the expected values. In other words, one could estimate a small model of how expectations are formed before estimating the basic model. Expectations are not rational in this case because they are not predictions from the basic model, but they are based on more information than merely the current and past values of one variable. An example of the use of this assumption is presented in Section 4.1.3. Although, as will be seen, this application was not successful, there is clearly room for more tests of this kind.

Other Unobserved Variables

In models in which disequilibrium is a possibility, there is sometimes a distinction between "unconstrained" and "constrained" (or "notional" and "actual") decisions. An unconstrained decision is one that an agent would make if there were no constraints on its decision variables other than the standard budget constraints. A constrained decision is one in which other constraints are imposed; it is also the actual decision. In the model in Chapter 3, for example, which does allow for the possibility of disequilibrium, a household may be constrained in how much it can work. A household's unconstrained consumption decision is the amount it would consume if the constraint were not binding, and the constrained decision is the amount it actually chooses to consume given the constraint. In models of this type the unconstrained decisions are observed only if the constraints are not binding, and so this is another example of the existence of unobserved variables. The treatment of these variables is a difficult problem in empirical work, and it is also a problem for which no standard procedure exists. The way in which the variables are handled in my model is discussed in Section 4.1.3.

2.2.3 Step 3: Specification of the Stochastic Equations

The next step is to specify the stochastic equations, that is, to write down the equations to be estimated. Since the stochastic equations are the key part of

any econometric model, this step is of crucial importance. If theory has not indicated the functional forms and lag lengths of the equations, a number of versions of each equation may be written down to be tried, the different versions corresponding to different functional forms and lag lengths. If the theoretical approach is the traditional one, theory has presumably chosen the LHS and RHS variables. The specification of the stochastic equations also relies on the treatment of the unobserved variables from step 2; the extra theorizing in step 2 also guides the choice of the RHS variables.

Theory generally has little to say about the stochastic features of the model, that is, about where and how the error terms enter the equations. The most common procedure is merely to add an error term to each stochastic equation. This is usually done regardless of the functional form of the equation. For example, the term $+ u_{it}$ would be added to equation i regardless of whether the equation were in linear or logarithmic form. If the equation is in log form, this treatment implies that the error term affects the level of the LHS variable multiplicatively. This somewhat cavalier treatment of error terms is generally done for convenience; it is another example of an unsatisfying aspect of the transition to econometric models, although it is probably not as serious as most of the other problems.

2.2.4 Step 4: Estimation

Once the equations of a model have been written down in a form that can be estimated, the next step is to estimate them. Much experimentation usually takes place at this step. Different functional forms and lag lengths are tried, and RHS variables are dropped if they have coefficient estimates of the wrong expected sign. Variables with coefficient estimates of the right sign may also be dropped if the estimates have t-statistics that are less than about two in absolute value, although practice varies on this.

If at this step things are not working out very well in the sense that very few significant coefficient estimates of the correct sign are being obtained, one may go back and rethink the theory or the transition from the theory to the estimated equations. This process may lead to new equations to try and perhaps to better results. This back-and-forth movement between theory and results can be an important part of the empirical work.

The initial estimation technique that is used is usually a limited information technique, such as 2SLS. These techniques have the advantage that one can experiment with a particular equation without worrying very much about the other equations in the model. Knowledge of the general features of the

other equations is used in the choice of the first-stage regressors (FSRs) for the 2SLS technique, for example, but one does not need to know the exact features of each equation when making this choice. If a full information technique is used, it is usually used at the end of the search process to estimate the final version of the model. If the full information estimates are quite different from the limited information ones, it may again be necessary to go back and rethink the theory and the transition. In particular, this may indicate that the version of the model that has been chosen by the limited information searching is seriously misspecified.

Sometimes ordinary least squares (OLS) is used in the searching process even though the model is simultaneous. This is a cheap but risky method. Because the OLS estimates are inconsistent, one may be led to a version of the model that is seriously misspecified. This problem presumably will be caught when a consistent limited information or full information technique is used, at which point one will be forced to go back and search using the consistent limited information technique. It seems better merely to begin with the latter in the first place and eliminate this potential problem. The extra cost involved in using, say, 2SLS over OLS is small.

2.2.5 Step 5: Testing and Analysis

The next step after the model has been estimated is to test and analyze it. This step, it seems to me, is the one that has been the most neglected in macroeconomic research. Procedures for testing and analyzing models are discussed in Chapters 7–10; they will not be discussed here except to note the two that have been most commonly used. First, the principal way that models have been tested in the past is by computing predicted values from deterministic simulations, where the accuracy of the predictions is usually examined by calculating root mean squared errors (Sections 8.2 and 8.3). Second, the main way that the properties of models have been examined is by computing multipliers from deterministic simulations (Section 9.2). As will be seen, both of these procedures, especially the first, are subject to criticism.

It may also be the case that things are not working out very well at this testing and analysis step. Poor fits may be obtained, and multipliers that seem (according to one's a priori views) too large or too small may also be obtained. This may also lead one to rethink the theory, the transition, or both, and perhaps to try alternative specifications. In other words, the back-and-forth movement between theory and results may occur at both the estimation and analysis steps.

2.2.6 General Remarks

The back-and-forth movement between theory and results may yield a model that fits the data well and seems on other grounds to be quite good, when it is in fact a poor approximation to the structure. If one searches hard enough, it is usually possible with macro time series data to come up with what seems to be a good model. The searching for models in this way is sometimes called "data mining" and sometimes "specification searches," depending on one's mood. A number of examples of this type of searching are presented in Chapter 4. Fortunately, there is a way of testing whether one has mined the data in an inappropriate way, which is to do outside sample tests. If a model is poorly specified, it should not fit well outside of the sample period for which it was estimated, even though it looks good within sample. It is thus possible to test for misspecification by examining outside sample results, and this is what the method in Chapter 8 does in testing for misspecification. (There is, however, a subtle form of data mining that even the method in Chapter 8 cannot account for. This is discussed in Section 8.4.5.)

Because of the dropping of variables with wrong signs and (possibly) the back-and-forth movement from multiplier results to theory, an econometric model is likely to have multiplier properties that are similar to what one expects from the theory. Therefore, the fact that an econometric model has properties that are consistent with the theory is in no way a confirmation of the model. Models must be tested using methods like the one in Chapter 8, not by examining the "reasonableness" of their multiplier properties.

It should also be emphasized that in many cases the data may not contain enough information to decide a particular issue. If, for example, tax rates have not been changed very much over the sample period, it may not be possible to discriminate between quite different hypotheses regarding the effects of tax rate changes on behavior. It may also be difficult to discriminate between different functional forms for an equation, such as linear versus logarithmic. In Chapter 4 a number of examples are presented of the inability to discriminate between alternative hypotheses. When this happens there is little that one can do about it except to wait for more data and be cautious about making policy recommendations that are sensitive to the different hypotheses.

2.3 Testing Theoretical Models

This is a good time to consider the second methodological question mentioned in Chapter 1, namely, what do econometric results have to say about

the validity of theories? It should be clear by now that transitions from theoretical models to econometric models are typically not very tight. It may be that more than one theoretical model is consistent with a given econometric model. If this is so, then finding out that an econometric model is, say, the best approximation among all econometric models is not necessarily a finding that a particular theory that is consistent with the model is valid. One may thus be forced to make weaker conclusions about theoretical models than about econometric models.

If it is possible to test the assumptions of a theoretical model directly, it may not be the case that one is forced to make weaker conclusions about theoretical models. The problem in macroeconomics is that very few assumptions seem capable of direct tests. Part of the problem is the aggregation; it is not really possible to test directly assumptions about, say, the way an entire sector chooses its decision variables. A related problem is that many macroeconomic assumptions pertain to the way in which agents interact with each other, and these assumptions are difficult to test in isolation. Assumptions about expectations are also difficult or impossible to test directly because expectations are generally not observed. Even if expectations were observed, however, it would not be possible to test the rational expectations assumption directly. In this case one needs a complete model to test the assumption. One is thus forced in macroeconomics to rely primarily on testing theories by testing econometric models that are derived (however loosely) from them. This procedure of testing theories by testing their implications rather than their assumptions is Friedman's view (1953) about the way theories should be tested. One does not, however, have to subscribe to Friedman's view about economic testing in general in order to believe that it holds for macroeconomics. Macroeconomic theories are tested indirectly not always out of choice, but out of necessity.

Given the indirect testing of theories and the sometimes loose transitions from theories to empirical specifications, it is not clear that one ought to talk in macroeconomics about theories being "true" or "false." Macroeconomics is not like physics, where on average theories are linked more closely to empirical tests. I have suggested (Fair 1974d) that it may be better in macroeconomics to talk about theories being "useful" or "not useful." A theory is useful if it aids in the specification of empirical relationships that one would not already have thought of from a simpler theory and that turn out to be good approximations. Otherwise, it is not useful. Although how one wants to label theories is a semantic question, the terms "useful" and "not useful" do highlight the fact that theories in macroeconomics are not as closely linked to empirical tests as are many theories in physics.

2.4 Expected Quality of Macroeconometric Models in the Long Run

An interesting question is how good one expects macroeconometric models to be in the long run, say in twenty or thirty years. It may be that behavior is so erratic and things like aggregation problems so severe that no model will be very good. This will show up in large estimated variances of prediction errors by the method in Chapter 8 and probably in large estimates of the degree of misspecification. Another way of stating this is that the structure of the economy may be too unstable or our potential ability to approximate closely a stable structure too poor to lead to accurate models. If this is true, models will never be of much use for policy purposes. They may be of limited use for short-run forecasting, but even here probably only in conjunction with subjective adjustments.

My research is obviously based on the premise that there is enough structural stability to warrant further work on trying to approximate the structure of the economy well. This is, of course, a premise that can only be verified or refuted in the long run, and there is little more that can be said about it now. It is interesting to note that the extensive use of subjective adjustments by the commercial model builders and their lack of much scientific research on the models may indicate lack of confidence in a stable structure.

It is also interesting to note, as mentioned in Chapter 1, that the lack of confidence in large-scale models has led to research on much smaller ones. In one sense this may be a reasonable reaction, and in another sense not. If the lack of confidence is a lack of confidence in a stable structure, the reaction does not seem sensible. It seems quite unlikely that the structure would be unstable in such a way as to lead small models to approximate it less poorly than large models. One should instead just give up the game and do something else. If, on the other hand, the lack of confidence in large-scale models is a feeling that they have gone in wrong directions, it may be sensible to back up for a while. In this case the premise is still that the structure is stable, and the issue is merely how best to proceed to try to approximate it well.

2.5 Nonlinear Optimization Algorithms

It may seem odd to put a section on nonlinear optimization algorithms in a chapter on macroeconomic methodology, but the solution of nonlinear optimization problems is an important feature of current macroeconomic research. In this book the following problems arise. (1) In the theoretical

model in Chapter 3 the decisions of the agents are based on the solutions of nonlinear multiperiod maximization problems. (2) The estimation techniques discussed in Chapter 6 require the solution of nonlinear optimization problems. (3) The optimal control problems discussed in Chapter 10 are set up as standard nonlinear maximization problems. (4) The estimation of rational expectations models discussed in Chapter 11 requires the solution of a nonlinear maximization problem.

For many nonlinear optimization problems, general-purpose algorithms are sufficient. One of the most commonly used is the Davidon-Fletcher-Powell (DFP) algorithm, which is discussed later in this section. For a number of problems, however, general-purpose algorithms do not work or do not work very well, and for these problems special-purpose algorithms must be written. As discussed in Section 6.5.2, the DFP algorithm does not seem to work for moderate to large FIML and 3SLS estimation problems. These problems must instead be solved using an algorithm designed particularly for them, the Parke algorithm. The other problems in this book for which special-purpose algorithms were written are the least absolute deviations (LAD) and two-stage least absolute deviations (2SLAD) estimation problems in Section 6.5.4 and the multiperiod maximization problems in Sections 3.1.2 and 3.1.3. The DFP algorithm does not work for the LAD and 2SLAD problems, and it was not tried for the multiperiod maximization problems because it seemed likely to be too expensive.

When general-purpose algorithms are used, it is not really necessary to know how they find the optimum as long as they do. They can, in other words, be treated as black boxes as long as things are going well. If the algorithms are not working well, knowledge of what they are trying to do may help either in modifying them for the particular problem or in designing new algorithms. In the remainder of this section a brief explanation of the DFP algorithm will be presented.

Consider the problem of minimizing $f(x)$ with respect to the elements of the $n \times 1$ vector $x = (x_1, x_2, \ldots, x_n)'$. (The problem of maximizing $f(x)$ is merely the problem of minimizing $-f(x)$.) The function f is assumed to be twice continuously differentiable. Approximating $f(x)$ by a second-order Taylor series about some point x^0 yields

$$f(x) \approx f(x^0) + g(x^0)'(x - x^0) + \frac{1}{2}(x - x^0)'G(x^0)(x - x^0), \tag{2.8}$$

where $g(x^0)$ is the $n \times 1$ vector of the gradient of $f(x)$ evaluated at x^0 and $G(x^0)$ is the $n \times n$ matrix of the second derivatives of $f(x)$ evaluated at x^0

Minimizing the RHS of (2.8) by setting the partial derivatives with respect to x equal to zero yields

$$g(x^0) + G(x^0)(x - x^0) = 0 \tag{2.9}$$

or

$$x = x^0 - [G(x^0)]^{-1} g(x^0). \tag{2.10}$$

Equation (2.10) forms the basis for many algorithms. Letting x^k denote the value of x on the kth iteration, one can iterate using (2.10):

$$x^k = x^{k-1} - [G(x^{k-1})]^{-1} g(x^{k-1}), \tag{2.11}$$

where some initial guess is used for x^0. If (2.11) is used exactly, the algorithm is called Newton's method, or Newton-Raphson's method. The matrix $[G(x^{k-1})]^{-1}$ is called the Hessian matrix.

Newton's method can be expensive because it requires calculating the Hessian matrix at each iteration, and much of the recent work in this area has been concerned with algorithms that do not require this calculation. The general formula for many of these algorithms can be written

$$x^k = x^{k-1} - \lambda^{k-1} H^{k-1} g(x^{k-1}), \tag{2.12}$$

where H^{k-1} is an $n \times n$ matrix and λ^{k-1} is a scalar. Algorithms based on (2.12) do two things at each iteration: (1) they choose a search direction $H^{k-1} g(x^{k-1})$, and (2) they choose a value for λ^{k-1} by carrying out a line search in this direction. (Newton's method is, of course, one of these algorithms, where $H^{k-1} = [G(x^{k-1})]^{-1}$ and $\lambda^{k-1} = 1$.) After the direction is chosen, the line search usually consists of fitting a second-degree polynomial to three points along the direction and then minimizing the resulting polynomial.

The algorithms differ in their choice of search directions. The DFP algorithm, which is of primary concern here, is a member of a class of methods called "matrix-updating" methods. Other names for this class include "quasi-Newton" and "variable metric." These methods never compute the Hessian, but instead build up an approximation to it during the iterative process by successive additions of low-rank matrices. The updating equation for the DFP algorithm is

$$H^0 = I, \tag{2.13}$$

$$H^{k-1} = H^{k-2} + \frac{\delta\delta'}{\delta'\gamma} - \frac{H^{k-2}\gamma(H^{k-2}\gamma)'}{\gamma' H^{k-2}\gamma}, \qquad k = 2, 3, \ldots,$$

where $\delta = x^{k-1} - x^{k-2}$ and $\gamma = g(x^{k-1}) - g(x^{k-2})$. There are a number of ways to motivate (2.13), but to do so here would take us too far afield; the interested reader is referred to Huang (1970) and Dennis and More (1977). (The original discussion of the DFP algorithm is contained in Davidon 1959 and Fletcher and Powell 1963.) It can be shown that if f is quadratic and if accurate line search is used, $H^n = G^{-1}$, where n is the dimension of x. Note that although algorithms like DFP do not require the computation of second derivatives, they do require the computation of first derivatives.

Another update that is sometimes used is

$$(2.14) \qquad H^0 = I,$$

$$H^{k-1} = H^{k-2} + \frac{\delta\delta'}{\delta'\gamma}\left(1 + \frac{\gamma' H^{k-2}\gamma}{\delta'\gamma}\right) - \frac{\delta\gamma' H^{k-2} + H^{k-2}\gamma\delta'}{\delta'\gamma}, \qquad k = 2, 3, \ldots,$$

where δ and γ are as above. This algorithm is called the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm. (See Dennis and More 1977 for references.) Once a program for the DFP algorithm has been written, the extra coding for the BFGS algorithm is small, and therefore many nonlinear optimization packages offer a choice of both the DFP and BFGS updating equations. My experience is that it generally does not make much difference which of the two updating equations is used. An example of the use of the two algorithms is reported in Section 10.4.

Another option that is sometimes available in nonlinear optimization packages is the method of steepest descent. This method simply uses $H^{k-1} = I$ for all k. It has very slow convergence properties, and it is not in general recommended.

The DFP algorithm has turned out to work well for many problems, and it is widely used. It does not, however, by any means dominate all other algorithms for all problems. There are also many problems for which it does not work in the sense that it does not find the optimum. My experience with the DFP algorithm is mixed but on the whole is fairly good. It has worked extremely well for the solution of optimal control problems, where in one case it was used to solve a problem of 239 unknowns (that is, $n = 239$). These results are reported in an earlier paper (Fair 1974a), where it can be seen that DFP easily dominated two other algorithms, one that required no derivatives (Powell's no-derivative algorithm; Powell 1964) and one that required both first and second derivatives (the quadratic hill-climbing algorithm of Gold-

feld, Quandt, and Trotter 1966). The solution of optimal control problems in this way is discussed in Section 10.2.

As noted earlier, DFP does not work for moderate to large FIML and 3SLS estimation problems, which seem to require special-purpose algorithms like the Parke algorithm. It also does not work for the minimization problem associated with the LAD and 2SLAD estimators. I have found it to work fairly well for the OLS or 2SLS estimation of a single equation that is nonlinear in coefficients.

My general strategy for dealing with nonlinear optimization problems is the following. If I choose to obtain and code analytic first derivatives, which is usually not the case, I merely solve the first-order conditions using the Gauss-Seidel technique (discussed in Section 7.2). In other words, I solve the equation system

$$g(x) = 0 \tag{2.15}$$

using Gauss-Seidel. I have had very good success with the Gauss-Seidel technique (with damping sometimes required), and the procedure of solving (2.15) avoids having to use any optimization algorithm. If first derivatives are instead computed numerically, then I usually begin with the DFP algorithm and only try other procedures if this does not work.

When first derivatives are computed numerically, they can be either "one-sided" or "two-sided." Consider the derivative of f with respect to x_1. One-sided derivatives are computed as $[f(x_1 + \epsilon, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_n)]/\epsilon$, where ϵ is a small number. Two-sided derivatives are computed as $[f(x_1 + \epsilon, x_2, \ldots, x_n) - f(x_1 - \epsilon, x_2, \ldots, x_n)]/2\epsilon$. Since $f(x_1, x_2, \ldots, x_n)$ is available at the time the derivatives are computed, one-sided derivatives require only one function evaluation per unknown, whereas two-sided derivatives require two. Both one-sided and two-sided derivatives were used for the results of solving the optimal control problems in Fair (1974a), and these results indicate that two-sided derivatives are not worth the extra cost. Little or no change in the number of iterations needed for convergence was obtained by the use of the two-sided derivatives. For the optimal control results in Chapter 10, on the other hand, slightly more accurate answers were obtained using two-sided derivatives, because the stopping criterion that was used for the Gauss-Seidel technique in solving the model was not small enough to allow highly accurate one-sided derivatives to be computed. This example is discussed in Section 10.4.

Note that the use of the DFP algorithm in conjunction with numerical

derivatives requires very little work to set up the problem. One merely needs to write a program (a subroutine when using FORTRAN) to compute f for a given value of x. Once this is done, the DFP algorithm merely calls this program many times in the iterative process. Each iteration requires n calls for the derivatives plus a few more for the line search. The calculations for each iteration other than the calculations involved in computing the function are generally very minor, so most of the computer time is taken in computing the function values. The estimates in Fair (1974a) for the one-sided derivative results show that this time is between 78 and 97 percent of the total time. For two-sided derivatives the percentages are even higher. It is thus important to code the function program efficiently. If numerical derivatives are used, it is easy to see why methods that require the calculation of second derivatives are likely to be expensive: $(n^2 + n)/2$ evaluations of the function are needed to calculate the second-derivative matrix, and for large n this is obviously expensive.

For purposes of the Fair-Parke program, I have coded the DFP and BFGS algorithms from scratch. The coding is straightforward except for the line search, which was coded as follows. (1) $\lambda = 1$ is tried. If this results in an improvement (a lower value of $f(x)$ than that of the previous iteration), $\lambda = 1.25$ is tried. If this results in an improvement, $\lambda = (1.25)^2$ is tried, and so on through $\lambda = (1.25)^9$. At the point of no improvement or at $\lambda = (1.25)^9$, a quadratic is fit to the three points $.8\lambda_s$, λ_s, and $1.2\lambda_s$, where λ_s is either the last value of λ that resulted in an improvement or $(1.25)^9$. The quadratic is minimized. The function is then evaluated for $\lambda = \lambda^*$, where λ^* is the minimizing value. A second quadratic is then fit to the three points $.95\lambda_{ss}$, λ_{ss}, and $1.05\lambda_{ss}$, where λ_{ss} is either $.8\lambda_s$, λ_s, $1.2\lambda_s$, or λ^*, depending on which one has yielded the smallest value of the function. This quadratic is minimized, and the function is evaluated for $\lambda = \lambda^{**}$, where λ^{**} is the minimizing value. The final value of λ is then taken to be $.95\lambda_{ss}$, λ_{ss}, $1.05\lambda_{ss}$, or λ^{**}, depending on which one yielded the smallest value of the function. (2) If $\lambda = 1$ does not result in an improvement, $\lambda = .5$ is tried. If this does not result in an improvement, $\lambda = (.5)^2$ is tried, and so on through $\lambda = (.5)^9$. At the point of improvement or at $\lambda = (.5)^9$, the quadratic fitting discussed in (1) is done.

The algorithm is stopped for one of five reasons: (1) no improvement is found for any value of λ tried at the current iteration; (2) the prescribed maximum number of iterations is reached; (3) the successive estimates of x are within some prescribed tolerance level; (4) at the current iteration the gradient values as a percentage of the respective x values are less than some

prescribed tolerance level in absolute value; or (5) the improvement in the function from one iteration to the next is within some prescribed tolerance level.

There is nothing subtle or sophisticated about this code, but it seems to work quite well for the types of problems I have dealt with. It may be that one could get by with fewer function evaluations for the line search (there is now a maximum of sixteen per iteration), but for problems with a large number of unknowns, these function evaluations are a small percentage of the function evaluations required to get the derivatives. With respect to the derivatives, the user has the option of deciding whether to use one-sided or two-sided derivatives and what step size to use.