WORKING PAPER SERIES

An Evaluation of Macroeconomic Models for Use at EIA

Vipin Arora

December 2013

Introduction

EIA has traditionally used macroeconomic models to produce forecasts and to evaluate the impact of different government policies. This document reviews the current EIA approach and alternatives from a methodological perspective. It begins with a short summary of different macroeconomic models and their strengths and weaknesses when used for policy analysis and in producing forecasts. This is followed by recommendations for possible use at EIA based on the capabilities of each model type. The mechanics of each specific macroeconomic model are reviewed next, along with additional details on policy analysis and forecasting. The final section is a technical appendix with the relevant mathematical detail on each model.

Summary and Recommendations

Overview of different macroeconomic models

EIA currently uses two macroeconometric models, the Global Insight U.S. macroeconomic model and the Oxford Economics global economic model. These are both simultaneous systems of equations based on standard macroeconomic theory which are estimated using historical data. Roughly speaking, the short-run structure of most macroeconometric models is based on Keynesian (IS-LM) theory so that variable outcomes are demand determined. Supply dominates in the long-run, which is based on the neo-classical growth model of Solow and Swan. While these are the overarching theories used in specification, macroeconometric models provide substantial flexibility in tailoring individual equations and variables to the needs of the modeler. The models can range from a few dozen equations to several thousand; they may be backward or forward looking; and macroeconometric models can be regional, national, or international in scope.

Computable general equilibrium (CGE) models are a different type of macroeconomic approach based on microeconomic theory. The decisions of consumers are based on the theory of consumer choice. Firms are profit maximizers, in that their objective is to make the difference between total revenues and total costs as large as possible. A general equilibrium is a situation where consumers and firms both do the best they can and all markets clear. This class of models is also flexible in its structure. They can be static; they may include a finite number of periods, or can extend to an infinite number of periods; and the level of detail in general equilibrium models can range from one consumer/one firm to dozens of different consumers with hundreds of sectors. Like macroeconometric models, CGE models can be regional, national, or international in scope.

Vector autoregressive (VAR) models are primarily statistical in nature and specify multiple macroeconomic variables in terms of other macroeconomic variables and error terms. One can think of a VAR model in terms of multiple equations, where each equation has a different endogenous variable on the left-hand side and the other endogenous variables and an error term on the right-hand side. The model variables can be chosen based on economic theory or other considerations, but are generally limited in nature (less than 10). VARs are estimated on historical data, and can be used to quantitatively and qualitatively understand the impact of different events or policies on macroeconomic variables. Such estimated VARs can also be used for short-term forecasting. Table 1 summarizes the key features of each model type.

Table 1: Comparison of macroeconomic models

	Structure	Theory	Expectations	Dynamics
Macroeconometric	Higher-level, non-optimization	IS-LM and neo-classical	Backward or forward-looking	Static or dynamic; stochastic or non-stochastic
General Equilibrium	Consumers/firms, optimization	Neo-classical	Backward or forward-looking	Static or dynamic; stochastic or non-stochastic
VAR	Non-optimization	Varies	Backward-looking	Dynamic; stochastic

Policy analysis and forecasting

Both the GI U.S. macroeconomic model and the Oxford GEM are used by EIA to analyze the impact of different government policies and to generate short and long-run macroeconomic forecasts. Each of the models outlined above can be used for policy analysis or forecasting, but have their relative strengths and weaknesses which are outlined below. Tables 2 and 3 summarize the major strengths and weaknesses of each type of model.

Policy analysis

The primary strength of macroeconometric models in policy analysis is the fact that they can evaluate a broad range of different policies. Because of their size, flexibility, and consistency with the national accounts macroeconometric models can evaluate both small and large changes to many different policy instruments. However, interpreting policy analysis is commonly believed to be problematic for this class of models. This is because changes in policy may invalidate some of the estimated relationships in the model that are based on historical data. Another concern is that the size of macroeconometric models can make them difficult to interpret.

The primary strength of general equilibrium models is that they can be used to evaluate a broad range of policies in terms of welfare. These models can also be used to track the flows of factors of production and goods in the economy in addition to their relative prices. The major deficiency of general equilibrium models for use in policy analysis lies in their limited representation of the financial sector. Another problem is that the numerical accuracy of such models may be poor, and choosing values for parameters can be problematic. Larger CGE models may also have intensive data requirements.

A very appealing aspect of using VARs for policy analysis is their flexibility. VARs are often the only technique available which can incorporate many variables without specifying them to be exogenous or imposing a strict model structure. One drawback of using VARs for policy analysis is that certain technical considerations require the modeler to make assumptions about interactions between variables which may be difficult to defend. Another objection to using VARs for policy analysis is the imposition of a linear structure combined with the assumption of errors that are random.

Table 2: Comparison of macroeconomic models in terms of policy analysis

Macroeconometric		
	Strengths	*Weaknesses*
1)	1. Broad range of policies.	1. Interpretation.
General Equilibrium		
	Strengths	*Weaknesses*
2)	1. Welfare measure.	1. Limited financial sector.
3)	2. Track flows between sectors.	2. Data requirements and parameter estimation.
VAR		
	Strengths	*Weaknesses*
4)	1. Flexible, can be atheoretical.	1. Need identifying assumptions.

Forecasting

Macroeconometric models are able to generate both short and long-term forecasts which are consistent between sectors and comparable with the national accounts. No other large scale macroeconomic model can do this. The weakness of macroeconometric model-based forecasting is the need to generate values for exogenous variables and the ad-hoc nature of add-factors. It can often be difficult to forecast future values of exogenous variables which are needed in the macroeconometric model. It is also unclear what role add-factors play in forecasting with macroeconometric models. There is no consensus on the interpretation of such factors when used.

Forecasting with CGE models is almost non-existent. The common view is that such models are ill-suited for generating accurate forecasts. VARs are routinely used for short-term forecasting purposes, and this is considered to be a strength of the VAR methodology. Evaluations of the forecasting performance of VAR models vary widely depending on what is being forecast and how the forecast is evaluated. VARs are much smaller than macroeconometric models, easier to estimate, and simpler to develop and maintain. Unlike macroeconometric models, VARs do not require one to make assumptions about exogenous variables when forecasting. One problem with forecasting with VARs is specification. The choice of variables to include in a VAR as well as how many lags to introduce can often be arbitrary. VAR forecasts are also difficult to interpret in economic terms, which is a strength of macroeconometric models.

Table 3: Comparison of macroeconomic models in terms of forecasting

Macroeconometric		
Strengths		*Weaknesses*
5)	1. Consistent short and long-term forecasts.	1. Need to generate values for exogenous variables.
6)	2. Add-factors allow for flexibility.	2. How to interpret add-factors?
General Equilibrium		
Strengths		*Weaknesses*
7)		1. Poor accuracy.
VAR		
Strengths		*Weaknesses*
8)	1. Small size and flexibility.	1. How to choose variables?
9)	2. Can be simple to develop.	2. How to interpret forecasts?

Recommendations

1. EIA should continue to use macroeconometric models as in the past.
2. A small CGE model should be developed at EIA.
    a. Initially this can be used with the oil price scenario project.
    b. If successful, the CGE model can be enlarged and then integrated with the WEPS+ modeling system for policy analysis purposes. The lessons learned from this process will provide a solid foundation in deciding whether or not to proceed in building a CGE model to use with NEMS for policy analysis.
3. Several VARs should be constructed and estimated by EIA.
    a. These can provide forecasts of different energy prices and production which are output from (or input to) the AEO and IEO. Such forecasts might also serve as a baseline for initial STEO or AEO runs.
    b. Other VARs can also be used to check and compare industrial and economic forecasts from either GI or Oxford Economics.

The Mechanics of Macroeconometric Models

Specification

Macroeconometric models are systems of stochastic (or behavioral) equations and identities. Stochastic equations are usually based at least in part on theory and are estimated using historical data, whereas identities are equations that hold by definition. These equations are comprised of both endogenous and exogenous variables. The values of endogenous variables are determined in the model; they are the solution to the model equations. Exogenous variables come from outside of the model structure.

The overall structure of macroeconometric models often follows a high-level approach. Major aggregates of interest are estimated using specific stochastic equations, and the remainder of variables are computed using identities or simple relations with the major aggregates. The definition of ``major aggregates'' differs between each model, and often depends on the interests of the modeler and purpose of the modeling exercise. Although all macroeconometric models begin with this basic framework, they will differ in the particular theory which is used in each stochastic equation, whether the equations are written in levels or in error correction form, the split between endogenous and exogenous variables, and in their size and scope.

Roughly speaking, the short-run structure of most macroeconometric models is based on Keynesian (IS-LM) theory so that variable outcomes are demand determined. The IS curve summarizes the relationship between the real interest rate and the level of income that arises in the market for goods and services. The LM curve plots the relationship between the real interest rate and the level of income that arises in the market for money balances. The intersection of these two curves gives the equilibrium real interest rate and income.

Most macroeconometric models imply a relationship between output and inflation in the short-run. That is, by increasing output (through government spending for example) a policymaker could get all of the benefits (such as reduced unemployment) at the cost of higher inflation. This relationship is known as the Phillips Curve. But this tradeoff only holds in the short-run in most of these models. As there is a transition to the long-run, the supply-side of the economy becomes more important and inflation becomes more and more the result of monetary factors.

Supply dominates in the long-run, which is based on the neo-classical growth model of Solow and Swan. The Solow-Swan theory isolates the key factors of economic growth in the long-run based on a one-sector production function. The key lessons that emerge are the importance of technology and savings in generating consistent and sustainable economic growth.

Estimation and testing

After a macroeconometric model is specified, estimation and testing are an important means of improving and verifying its performance. Once the model has been specified, the next step is to incorporate historical data and estimate the relevant coefficients. Most macroeconometric models are estimated equation-by-equation using OLS. However, the structure of the model and underlying data may need to be changed in order to make this a plausible estimation strategy.

Macroeconometric models are systems of simultaneous equations. One problem that often arises in estimating such systems is simultaneous causality between the variables on the left and right-hand sides of each equation. This can lead to violations of OLS assumptions, and means that the OLS estimates of equations are biased and inconsistent. One way to circumvent this issue is to use two-stage least squares estimation. However, modelers often ignore simultaneity issues, arguing that the bias is not large enough to seriously impair results.

Another issue which affects estimation is that many macroeconomic variables are growing over time. Using OLS for estimation with such trending variables leads to biased test statistics which may invalidate standard hypothesis testing. There are several ways to overcome this issue. A simple approach is to filter or difference the data before estimation. Differencing variables for use with an error-correction model is another way around stationarity issues. This requires the variables in each equation to be cointegrated. A final option is to ignore the stationarity issues, or to add a deterministic trend to the relevant equations. Neither of these approaches validates test statistics estimated on trending variables, but the use of such statistics can be tested through a bootstrapping procedure for each equation.

Macroeconometric models are commonly tested during and after specification, estimation, and solution. In general, single equation tests are performed throughout the model building and estimation stages, while full model tests are performed after the model has been solved. A common type of single equation test is to add a variable or a set of variables to a particular equation and test for statistical significance. This can take the form of adding a lag or multiple lags (or leads) of endogenous and exogenous variables and testing for statistical significance (t-test), testing for joint significance (F-test), or both. A deterministic time trend can also be added and checked for statistical significance, as can an intercept.

Other single equation tests check for different violations of estimation assumptions. One common test is to look for serial correlation of the error term in individual equations. Heteroscedasticity is another violation of such assumptions which is often tested, as is the normality of errors in each equation. If detected, any of these errors can be fixed relatively easily by adding lags or additional regressors, or by using corrected standard errors. Some modelers also check individual equations for parameter stability, or structural breaks. There are different variants of such tests which can be employed. However, once parameter instability is found it is unclear how such a problem should be addressed.

The simplest and most common way to test full models is to see how well their forecasts fit the data. The evaluation of forecasts can be either in-sample or out-of-sample. In-sample evaluation consists of using the first t observations of a sample to compute the respective coefficients. The modeler can then calculate the error for the remainder of observations in the sample, say t+50. Out-of-sample forecast errors are calculated with observations which are not currently in the sample, and usually become available only over time.

Whichever method is used, the forecaster must pick a way to summarize the forecast errors. A popular choice to summarize forecast accuracy is mean absolute deviation (MAD). This is the average of the absolute values of the forecast errors. This is sometimes also called mean forecast error (MAE), and is appropriate when the cost of forecast errors is proportional to the absolute size of the forecast error. It

is sensitive to scaling. Root mean square error (RMSE) is also sensitive to scaling, and is the square root of the average of the squared values of the forecast errors. This measure weights large forecast errors more heavily than small ones.

Mean absolute percentage error (MAPE) is not sensitive to scaling, and is the average of the absolute values of the percentage errors. It is appropriate when the cost of the forecast error is more closely related to the percentage error than to the numerical size of the error. Another measure is the correlations of forecasts with actual values. The percentage of turning points criteria is a 0/1 measure which summarizes if turning points were forecast correctly. Each of these has its variations, and there are other methods as well which may be used for specialty forecasts.

Policy analysis

Macroeconometric models are commonly used at EIA and in other institutions for policy analysis. This is generally done by changing an exogenous model variable and comparing simulation results from a given baseline. Compared to both CGE and VAR models, using macroeconometric models for policy analysis is relatively easy because of their design.

Strengths and Weaknesses

The primary strength of macroeconometric models in policy analysis is the fact that they can evaluate a broad range of different policies. Because of their size, flexibility, and consistency with the national accounts macroeconometric models can evaluate both small and large changes to many different policy instruments. No other class of macroeconomic models can match their detail and consistency with the national accounts.

However, interpreting policy analysis because of the handling of expectations can be a weakness for this class of models. Policy shifts may change consumer and firm expectations, and possibly behavior. The difficulty in dealing with these expectational (or behavioral) changes is that past responses may not be a reliable guide to future responses. In macroeconometric models, various coefficients summarize the responses of variables to one another, and the coefficients are estimated based on historical data. The standard argument is that a policy shift invalidates these coefficient values because it inherently changes the relationships between the variables, through either consumer or firm expectational or behavioral changes.

Another concern is that the size of macroeconometric models can make them difficult to interpret. It can be unclear what is driving different forecasts of model results. Their top-down approach also is unable to account for interactions between different sectors, as there are no explicit links specifying the input of one sector might be the output of another. Rather, changes to each variable are determined in the stochastic equations at a relatively high-level, and make their way into more detail from this higher level.

Forecasting

The primary use of macroeconometric models at EIA is for forecasting. Once the model has been specified, estimated, and tested it can be used to predict the future values of endogenous variables. The inputs required are forecasts of exogenous variables over the sample period, and the values of any add-

factors if they are being used to generate the forecasts. Most commercial macroeconometric models arrive with a baseline forecast over a particular time horizon, along with alternative forecasts.

Strengths and weaknesses

Macroeconometric models are able to generate both short and long-term macroeconomic forecasts which are consistent between sectors and comparable with the national accounts. VARs may be able to forecast individual variables more accurately, but are unable to guarantee consistency between sectors.

The weakness of macroeconometric model-based forecasting is the need to generate values for exogenous variables and the ad-hoc nature of add-factors. It can often be difficult to forecast future values of exogenous variables which are needed in the macroeconometric model, although this can be done by using VARs. It is also unclear what role add-factors play in forecasting with macroeconometric models. There is no consensus on the interpretation of such factors when used.

The Mechanics of General Equilibrium Models

Basics

The general equilibrium approach to macroeconomics seeks to explain the aggregate economy based on microeconomic theory. In its most basic form, this is done by specifying the objectives and constraints of both consumers and firms and then closing the model by assuming there is equilibrium. Consumers are assumed to maximize utility subject to their individual budget constraints. Perfectly competitive firms are assumed to maximize profits (or minimize costs) subject to their individual production technology. A general equilibrium is a situation where consumers are maximizing utility, firms are maximizing profits, and all markets clear.

The decisions of consumers are based on the theory of consumer choice. Consumers determine what is feasible, what is desirable, and then choose the most desirable from the feasible. The feasible choices for a consumer depend on their income, and this is summarized in a budget constraint. The preferences of a consumer are captured by a utility function, which is defined over goods or activities. The process of choosing the most desirable from the feasible is also called optimization or maximization, whereby consumers choose the consumption of goods or activities to maximize utility.

Firms are profit maximizers, in that their objective is to make the difference between total revenues and total costs as large as possible. Total revenues are the number of goods sold times their price, and total costs are those associated with any inputs to production. These inputs can be factors of production, such as capital and labor, or intermediate inputs such as energy. The firm's choice is constrained by the available production technology, summarized in the form of a production function. Although firms can have market power, the majority of general equilibrium models use a perfectly competitive environment.

A general equilibrium is a clearly defined concept in this framework that rules out many alternative scenarios, but ensures that any results are consistent with the underlying structure of the model. It is a situation where consumers and firms both make optimal decisions and all markets clear. If firms are perfectly competitive and both consumers and firms take all prices as given, this is also referred to as a competitive equilibrium.

Apart from these fundamentals, general equilibrium models can differ among many dimensions. These models can be static, they may include a finite number of periods, or can extend to an infinite number of periods. It is possible to introduce government expenditures and taxes into general equilibrium models as well. Although most of these models are specified in real terms, money and a monetary authority can be incorporated. The level of detail in general equilibrium models can range from one consumer/one firm to dozens of different consumers with hundreds of sectors.

Types of general equilibrium models

There are two broad categories of general equilibrium models. Computable general equilibrium (CGE) models are larger representations of this class which are primarily used for policy analysis. Dynamic stochastic general equilibrium models (DSGE) are smaller and more widely used in research activities. While similar in basic structure, these can be very different in implementation.

CGE models are used in many government departments for different types of policy analysis.[1] These models can be static, they may have a small number of future periods, or they may extend many years into the future. Their level of sectoral detail depends on the purposes for which they are built, so there is no standard CGE model for policy analysis. Uncertainty is not explicitly incorporated into CGE models, and they are not used for forecasting purposes.

Dynamic CGE models can use either forward or backward looking expectations in determining the future values of variables in simulations. Examples of backward looking expectations are myopic and adaptive expectations. When expectations are myopic, consumers expect the future value of a variable to be the same as a past value. This future value is expected to be a weighted average of past values when expectations are adaptive. In contrast, perfect foresight is assumed when forward looking expectations are used in these models. This means the consumer is perfectly able to predict the future values of any variables of interest.

DSGE models do explicitly incorporate uncertainty and are predominantly forward looking. These models use rational expectations, which imply that consumers are correct on average in forming their expectations about the future values of variables. DSGE models cannot be made very large due to the incorporation of uncertainty, and this limits their usefulness in detailed policy analysis. Their primary uses to date have been in the research work at universities and central banks. Some recent progress has been made in using DSGE models to forecast different macroeconomic variables, but this is an emerging research area.

Policy analysis

The primary use of both CGE and DSGE models is for policy analysis. Given their different dynamic structures, variations in the way expectations are formed, and integration of uncertainty, the ways in which they are used for policy analysis differ. Understanding these uses requires some additional information on the way in which data is incorporated into each type of model.

CGE models incorporate data through a calibration process. There are many underlying parameters in general equilibrium optimization problems which can take different values. These so-called structural parameters (because they are assumed to be unchanging over time) must be assigned values by the modeler before any quantitative simulation results can be generated. The parameters generally stay the same value throughout the simulation, and help to quantify production possibilities, welfare, different aspects of trade, and consumption.

The standard approach is for the modeler to first choose a base year for the data. This base year data is used to construct a social accounting matrix (SAM). This matrix is a record of transactions that take place within an economy during a given year. Specifically, it is an organized matrix representation of all transactions and transfers between different production activities, factors of production, and institutions (households, corporate sector, and government) within the economy and with respect to the rest of the world. A SAM is represented by a square matrix which lists these categories on the

[1] In the U.S. they have been used by the Environmental Protection Agency (EPA), the Department of Energy (DOE), the International Trade Commission (ITC), the Congressional Budget Office (CBO), the Department of Homeland Security (DHS), and the Department of Agriculture.

vertical and horizontal axes. The sums of each column in a SAM must match the sums of each corresponding row. Using the data from this SAM, parameter values in the CGE model equations are chosen so that the model is able to reproduce the data in the SAM. That is, the CGE model is able to solve for the base year data given these parameter values. This is often called the baseline database.

The baseline database in a dynamic CGE model includes the values of variables in the base year and all other years relevant to the simulation. The parameter values will be based on the SAM in the base year as in the static case. If the dynamic CGE model is backward-looking, a baseline database can be built by using the model equations. Given the base year data and parameter values, the model can generate values for all other variables for the entire simulation period.

A dynamic CGE model with forward looking expectations must be handled differently because the equations contain the future values of variables. No solution is possible without specifying a value for each of these forward-looking variables. There are three ways to calibrate such a model which exploit the perfect foresight assumption in expectations. The first is to choose the parameter values so that the data does not change from that in the base year. This is called putting the model at steady state. The values of any future variables in the equations are the same as the base year value.

It is more common to assume that all variables grow at some exogenously specified rate. If the rate is the same for all variables in the model, this is called putting the model on a balanced growth path. The values of any future variables in the equations are the same as the base year value times this growth rate. The third option is to specify that certain variables match exogenously given growth rates. The most common is to specify GDP growth rates, population growth rates, and the growth rates of certain types of energy consumption. The future values of the pre-specified variables then equal the previous year's value times the growth rate in the current year for that particular variable. The growth rates of variables which are not pre-specified can differ.

Once this baseline database is available a policy simulation can be conducted. Usually, this takes the form of 'what if' analysis. Different types of policies are evaluated by comparing the deviations of certain variables from their baseline values. The results are point estimates conditional on the initial baseline database and parameter values. Sensitivity analysis can be conducted by varying the baseline database or changing certain parameter values and then conducting the policy simulation again.

DSGE models are both calibrated and estimated. When calibrated, these models do not use a base year to get parameters values. Rather, some parameters are specified using commonly accepted values, and others are chosen so that certain ratios in the model match long-run averages in the data. For example, it is standard to choose the depreciation rate of capital in order to match a long-run capital to output ratio. DSGE parameter values can also be estimated, in which case a sample of historical data consistent with the model yields these parameter estimates.

In either case there is no baseline database for a DSGE model. Simulations from these models evaluate the impact of exogenously specified 'shocks'. These shocks are unexpected movements in certain variables, and examples include shocks to technology in production, tax rates, utility, and money supply, among others. Because the shocks are stochastic, their impact is not evaluated through point estimates, as any such estimate will depend on the particular realization of the shock.

The most common approach is to compute average statistics over many simulations. These statistics can include the correlations between certain variables, their standard deviations, as well as various autocorrelations. Alternatively, DSGE models produce impulse response functions which show the response of model variables to a one-time shock. These are used more for qualitative analysis, to gauge the magnitude and response of the variables to certain shocks.

Policy simulations in DSGE models are conducted in several ways. One way is to compare the results of different models with and without a policy change. For example, one could evaluate the long-run properties of a model with and without a particular tax. These are also called counterfactual simulations. These policy changes can also be traced through the model using the impulse response functions. Another way to use DSGE models for policy analysis is in quantifying the relative size of distortions in terms of welfare. This is common with taxes, but increasingly in quantifying the costs of regulation as well.

Strengths and weaknesses

The primary strength of general equilibrium models is that they can be used to evaluate a broad range of policies in terms of welfare. Because of their strong theoretical foundations, such models are also explicit about model assumptions, and the mechanisms driving results can be related to the consumers and firms in the model. This means that the behavioral responses in the model are not assumed but derived from the underlying structure of the model.

General equilibrium models are also able to overcome theoretical issues related to expectational change in the economy. Because the expectations of consumers and firms change with policy shifts, models which are based on assumptions about past behavioral relationships may not be able to capture these shifts. General equilibrium models avoid this issue because the behavior of consumers and firms is modeled directly, so that the responses to policy shifts are incorporated.[2] A final benefit of using general equilibrium models in policy analysis is that they can be used to track the flows of factors of production and goods in the economy in addition to their relative prices. The design of these models is such that all of the flows must be consistent to make income equal to expenditures in the economy.

The major deficiency of general equilibrium models for use in policy analysis lies in their limited representation of the financial sector. This makes them ill-suited for studying monetary policy or financial issue more generally. Although incorporating technological change is theoretically consistent with the general equilibrium structure, it can often be problematic to do so when solving and simulating such models. The production structure of CGE models is also top-down, meaning the level of detail in a particular industry may be less than desirable.

Another problem is that the numerical accuracy of such models may be poor due to the explicit specification of the environment and optimizations. Choosing values for parameters can be problematic in larger models, and these larger models also have intensive data requirements. In addition to these requirements, the base year chosen for the data can influence the simulation results substantially. Finally, the dynamics of CGE models are not usually suited for short-run analysis. This is because

[2]This is referred to as the weak-form of the Lucas critique. The strong form says that not only consumer and firm expectations, but also their behavior changes with policy shifts.

standard representations allow the perfect mobility of factors of production and may not incorporate unemployment explicitly.

Forecasting

Although both CGE and DSGE models are capable of forecasting, only DSGE models are used for this purpose in practice. In theory, the equations from a backward-looking CGE model provide forecasts of each variable in the model after the base year. However, it is well-known that these projected values are highly inaccurate in future periods. CGE models with forward-looking expectations can also generate forecasts, but this is not common both because they are likely to be inaccurate and solving such models is difficult.

The problem is that the model has yet to be solved for the future periods, making specification of the correct future value problematic. In order to forecast in such a model an iterative procedure must be used. The modeler makes a guess for the values of forward-looking variables in future periods, solves the model to get the implied values, and uses these implied values as the new starting point. The procedure continues until the difference in the implied values between successive iterations is below some threshold.

DSGE models have recently become more widely-used for forecasting. These models can be solved to yield a representation which generates the future value of each variable in the model in terms of past and current values of other variables and shocks. Given parameter values and the values of any exogenous variables the models can then be used for forecasting purposes.

Strengths and Weaknesses

DSGE model-based forecasts have been shown to be comparable in terms of their accuracy to certain macroeconometric models for a small number of variables. However, it is still unclear how these different types of models compare in their fit with respect to a broader set of variables. A particular issue in forecasting with DSGE models is that it can be very technically demanding, which may limit is appeal. DSGE models are also much smaller than commercial macroeconometric models, which limits detailed forecasting.

Other issues and assumptions

Because of the large number of individuals and goods in actual economies, two simplifications are often made to the consumer's utility maximization problem in general equilibrium models. First, it is common to aggregate different goods into composites. There can be one composite consumption good that aggregates all of the goods in the economy for consumption, or there can be several. The other simplification is to assume that there are one or more ‘representative’ consumers in the economy. Using this short-cut allows analysis in terms of only a few consumers (at most), because all consumers in each group represent a particular type, for example high and low-income individuals.

The same types of assumptions are made on the firm side as well. The standard representative firm stands in for all the firms in an economy or a particular sector. Implicit in this assumption is that the size of the firm is irrelevant for the firm’s marginal cost, so there is no advantage to either a small or large firm. Variations to this assumption are common, however, but not incorporated into standard models.

The Mechanics of Vector Autoregressions

Basics

An autoregressive (AR) model is one which specifies that the current value of a variable depends only on its own past values and the current value of an error. The number of past values (lags) of a variable specified in an AR model is generally designated by the letter p. An AR(p) is one where the current value of a variable is specified in terms of p previous values and the current error term.

The error term plays an important role in time series analysis. In an AR model it is often assumed to be independent and identically and normally distributed with mean zero. Independence refers to the fact that the value of the error at any point in time has no bearing on its value at another time. Identical and normal distribution with zero mean says that each time an error term is observed the probability it will take a certain value is given by the normal distribution, and on average this is expected to be zero. The coefficient values of an AR model can be estimated in a standard way by using ordinary least squares (OLS).

A natural extension is to write the current value of the variable of interest (the dependent variable) in terms of its own lags and the current values and lags of other variables. To estimate this larger equation with OLS requires assuming that the new variables are independent, that they are not impacted by changes in the dependent variable. Such independent variables are referred to as exogenous, while the dependent variable can also be called endogenous. In practice it is very difficult to find true independent variables in the macroeconomic context. Most prices and quantities of interest such as interest rates, money supply, GDP, price indexes, and exchange rates likely impact each other.

Extending an AR model to a vector autoregression (VAR) can overcome this problem. These models specify multiple endogenous variables in terms of other endogenous variables and error terms. One can think of a VAR model in terms of multiple equations, where each equation has a different endogenous variable on the left-hand side and the other endogenous variables and an error term on the right-hand side.

The structural form of a VAR model specifies each endogenous variable in terms of the current and past values of all endogenous variables and the current value of all error terms. The error terms in the structural form are called innovations or structural shocks, and are assumed to be independent, identically and normally distributed with mean zero. Because the structural form has both current and lagged values of endogenous variables on the right-hand side, OLS cannot be used to estimate the system. To use OLS, the structural form can be rearranged to the reduced form, which has the current value of each endogenous variable only in terms of lagged values of the endogenous variables and a reduced form error term. This error term is not the structural shock, but due to the rearrangement is some function of multiple structural shocks. The reduced form can be estimated one equation at a time by OLS, which yields the coefficient values associated with each lagged endogenous variable.

General issues and assumptions

As with any estimation using OLS, the relationship between the left-hand and right-hand side variables is assumed to be linear, and all OLS requirements are assumed to hold. One particular issue is that the

reduced form errors are correlated across equations in a VAR. If each equation has the same variables on the right-hand side, OLS is still the best estimation method. This is the primary reason why many VARs have the same variables and lags in each equation.

There is the further question of how the endogenous variables are selected for inclusion. The usual procedure is to refer to a relevant economic model, but this may not always be possible. As with any empirical model, the user seeks to incorporate the fewest number of variables in hopes of keeping it as simple as possible. Even with parsimonious models, one issue that often arises in the case of VARs is that the coefficient estimates are statistically insignificant. This is called overparameterization, and occurs because there may be a large number of coefficients to be estimated in a VAR: each endogenous variable has an equation which has lags of every other endogenous variable.

Choosing the appropriate lag length can be more difficult, although there are a variety of tests which can aid in the process. The most common way to proceed in order to preserve the symmetry of the system is to choose the same lag length for all variables in the system. It is also generally accepted when working with macroeconomic variables that the lag-length should be at least one year to capture any seasonality in the data, although some modelers do not follow this practice. The tests of lag length amount to finding the lag length of the model which provides the best fit according to some criterion.

There is also considerable debate on how to proceed when variables in a VAR are trending over time, which impacts the results from OLS estimation. If the goal is to uncover the interrelationships between the variables (as in policy analysis), then the general view is to leave the variables untransformed when estimating the VAR. While the accuracy of the parameter coefficients will be suspect in this case, transforming the data can throw away information on their interrelationships. This advice does not hold if the goal is to use a VAR for forecasting, in which case the data must be transformed in some way to remove the trend.

Policy analysis

VARs are used for policy analysis because they can assess the quantitative and qualitative relationships between endogenous variables. The basic idea is that an unexpected change in one endogenous variable due to a structural shock alters the current value of that variable, which then may alter all other endogenous variables in the system. Because the structural shock is exogenous, in the sense that is it unpredictable and uncorrelated with anything else in the system, the resulting movements in other endogenous variables are interpreted as due to the original shock.

Using a VAR for policy analysis first requires recovering the shocks from the reduced form estimation. Because of the transformation from the structural form to the reduced form, there are not enough variables in the reduced form to uniquely determine the values of the variables in the structural form. To get around this, the modeler must restrict the value of a certain number of variables in the structural system before using the VAR for policy analysis. This is called identification of the system.

The simplest and most common identification technique is called recursive identification. When using recursive identification the modeler specifies the order in which the structural shocks impact endogenous variables in the current period. The variable ordered first can impact all other variables in the VAR during the current time period. This means that the structural shock in the first equation can

impact all of the other variables in the system. The variable ordered second impacts only the variables ordered after it, but not the variable ordered first. This means that the structural shock in the second equation impacts all variables except the variable ordered first. This says that unexpected movements in the second variable, by assumption, cannot change the current value of the first variable. The variable ordered third cannot change the value of the first two in the current period, but can alter the others. The pattern continues for the remainder of variables in the VAR system.

Recursive identification is the simplest identification method, but it imposes very strong restrictions on the VAR system. At a minimum, the ordering chosen for the variables should be based on economic theory. Often, economic theory is unable to justify such strong restrictions on endogenous variables so other methods have been developed. Structural identification is similar, but alters the symmetry of the recursive ordering. In this case economic theory is used to specify which endogenous variables might impact other variables in the current period. The flexibility of this approach is that all causal links in the system can be specified (as with the recursive ordering) or just one for a single relationship. The number of identifying restrictions used will depend on the theory.

Long-run restrictions can also be used to identify the structural shocks. These are more commonly used for smaller VARs. Recently, VARs identified by sign-restrictions have been widely-used. In this case, the sign of the response of one endogenous variable to a shock in another is specified. This makes it much easier to apply economic theory to the identification procedure. Because this is also a weaker restriction, there are some technical issues relating to the uniqueness and robustness of the results from such models which is currently an area of research.

Irrespective of how VARs are identified in policy analysis, the results are quantified through a procedure known as innovation accounting. Innovation accounting is comprised of impulse response analysis, variance decompositions, and historical decompositions. These three tools give different ways to assess the impacts of structural innovations in one variable on the other endogenous variables.

An impulse response function is a powerful way to summarize the instantaneous and continuing impact of movements in structural innovations on the endogenous variables in a VAR. It graphically summarizes the impact of a one time, one unit increase of a structural innovation on each endogenous variable in the system. The modeler can isolate the direction and magnitude of the responses of each endogenous variable to each structural innovation using impulse response analysis.

Variance decompositions, technically forecast error variance decompositions, quantify the extent to which each particular structural innovation contributes to the total forecast error of an endogenous variable. The forecast error is the difference between the actual value and value predicted by the VAR. A variance decomposition allows the modeler to understand which particular structural innovation was most important for observed forecast error in a particular endogenous variable. This might indicate the relative importance of certain structural innovations in the movements of that endogenous variable.

Historical decompositions are similar to variance decompositions, but look backwards. They decompose the value of an endogenous variable in the VAR at any point in time terms of only the structural innovations. This allows the analyst to make statements about the relative historical importance of one structural innovation versus another in determining the value of an endogenous variable.

Strengths and weaknesses

VARs are widely used for policy analysis in research institutions, and there are pros and cons to this approach. A very appealing aspect of using VARs for policy analysis is the flexibility of the procedure. There is substantial freedom for the modeler to add or drop variables, add or drop lags, and vary time periods as required. The technique also has wide policy applicability, and has been used in contexts relating to monetary policy, fiscal policy, economic uncertainty, factors impacting the oil price, and many others. VARs are often the only econometric technique available which can incorporate many variables without specifying them to be exogenous or imposing a strict model structure. Impulse responses, variance decompositions, and historical decompositions are also very powerful tools for analysis.

The drawbacks of VARs boil down to identifying restrictions. Often these are ad-hoc and may be hard to justify based on any type of theory. This is particularly true of recursive identification, but can also true of alternative techniques. Where economic theory can be used, as with sign-restricted VARs, there is often a problem of finding a unique solution, or at least the most likely solution. Another objection to using VARs for policy analysis is the imposition of a linear structure combined with the assumption that structural innovations exist. Are there really structural innovations? Ultimately this question is difficult (maybe impossible) to answer and depends on the beliefs of the modeler. But it explains some of the skepticism which VARs have received when used for policy analysis.

A final point regarding the use of VARs in policy analysis is their interpretation. By construction VARs provide no mechanism to relate the identified quantitative and qualitative impacts. To some this is their biggest strength because it allegedly lets the data speak. To others this is their biggest weakness because the responses and impacts have no story.

Forecasting

VARs are routinely used for short-run forecasting purposes. The basic procedure is to first estimate the coefficient values from the reduced form VAR as described above. Using these coefficient values and the current values of each endogenous variable, the VAR model can forecast an arbitrary number of periods into the future. For example, in the AR(1) case, the value of the variable next period is expected to be the value of the variable in the current period times the estimated coefficient value. The expected value in two periods depends on the expected value next period, which depends on the current value and coefficient estimate. In this way a VAR can forecast into the future using only current values and coefficient estimates.

This also highlights why the estimated coefficient values are so important when using a VAR for forecasting. The fact that the coefficients may be suspect because the data is trending or the VAR is overparameterized is a big concern when generating forecasts. If some of the endogenous variables have a trend, the forecaster can either modify the data or modify the VAR model. The data can be modified in various ways, all of which work to remove the trend. The problem is that these methods also change the properties of the data, and should be avoided if at all possible. One situation where such transformation can be avoided is when it is known beforehand that some of the trending variables have a long-run relationship that does not have a trend. In this case the long-run relationship can be incorporated by modifying the VAR model.

A common approach to overcome issues related to overparameterization is to throw out the insignificant coefficients and re-estimate the VAR. This might mean discarding variables entirely, or just certain lags of the variables. This is termed a near-VAR. In larger VARs it can be difficult to decide which variables or lags of variables in each equation should be discarded. An additional complication arises when equations in the VAR contain different variables or lags of variables. A modification of the OLS procedure must be used in this case.

Bayesian VARs are an alternative to near-VARs that do not require the forecaster to take an all-or-nothing approach on the values of coefficients. That is, dropping a variable or a lag of a variable effectively sets its value to zero. This can be avoided with Bayesian methods by placing a prior probability on the value of each coefficient. In this way, fuzzy restrictions can be placed on the coefficients, which may be over-ridden by the data. However, using Bayesian methods adds computational complexity to the estimation.

Strengths and Weaknesses

Evaluations of the forecasting performance of VAR models vary widely depending on what is being forecast and how the forecast is evaluated. As a general rule of thumb, VAR forecasts may not be as good as AR forecasts for individual variables. However, their accuracy varies dramatically with the number of lags used, and Bayesian VARs in general are comparable in performance to AR models.

The strengths of VARs in forecasting come from their flexibility when compared to alternatives. Although they are larger than AR models, VARs are able to incorporate many additional variables. They are much smaller than macroeconometric models, easier to estimate, and simpler to develop and maintain. Unlike macroeconometric models, VARs do not necessarily require one to make assumptions about exogenous variables when forecasting. The problems with their use begin boil down to specification. The choice of variables to include in a VAR as well as how many lags to introduce can often be arbitrary, especially when using a near-VAR. VARs are also larger than AR models, and can be more complex to estimate. Finally, they are difficult to interpret in economic terms, which is a strength of macroeconometric models.

Appendix A: The Mechanics of General Equilibrium Models

Overview

This document summarizes some basic technical details of general equilibrium models.[1] It begins with a static general equilibrium model, characterizes the individual components of such models, and describes the equilibrium solution and some simple experiments. A small CGE model is also put forth in this section, along with a description of the construction and uses of social accounting matrices. The second section extends the static model to the two period case to introduce dynamics. Again, the individual components are outlined and then brought together to show the equilibrium solution using two slightly different models. Classical theoretical results such as the permanent income hypothesis and Ricardian Equivalence are also highlighted in the context of the models. The final section extends to the fully dynamic case, with a similar characterization of each individual component. This section also introduces issues related to expectations, uncertainty, and gives popular examples of fully dynamic models.[2]

Static Models

Static models, those which do not incorporate a time dimension, are the most basic general equilibrium models. They are often used to study long-run issues. This section builds a simple, static general equilibrium model step-by-step to illustrate its mechanics. The first section below describes the consumers utility maximization problem. The second details the profit maximization problem of the firm. These two parts are brought together in the next section, where the general equilibrium concept is explained. A slightly larger computable general equilibrium (CGE) model is put forth after this as an example. The final section outlines the construction of a social accounting matrix (SAM), which is used to construct a baseline scenario for CGE models. The majority of this section is based on Williamson (2011).

[1] None of the material covered here is original. As much as possible, the original sources of the equations, explanations, and examples have been cited.

[2] Much of the discussion is based on Williamson (2011) and Wickens (2008), see either for additional details.

Consumers

The decisions of consumers are based on the theory of consumer choice. Consumer's determine what is feasible, what is desirable, and then choose the most desirable from the feasible. The feasible choices for a consumer depend on their income, and this is summarized in a budget constraint. The preferences of a consumer are captured by a utility function, $U(\cdot)$, which is defined over goods or activities. The process of choosing the most desirable from the feasible is also called optimization or maximization, in that consumers choose consumption of goods or activities to maximize utility. The appendix specifies the details of this decision making process in the general case, building up from preference relations.

Because of the large number of individuals and goods in actual economies, two simplifications are often made to the consumer's utility maximization problem. First, it is common to assume that there are only two goods which consumers desire. The first good is a physical good for consumption purposes, or a consumption good. This may be thought of as a composite of all goods available in an economy. The second good is leisure, which is time not spent working. The inclusion of leisure may seem strange, but it can include recreational activities, sleep, or unpaid work at home. Consumers derive utility based on their consumption of a bundle comprised of physical goods and leisure, (C, l). The utility function is a numerical representation of the consumer's preferences, in this case $U = U(C, l)$. When comparing two different consumption bundles, (C_1, l_1) and (C_2, l_2), a higher value of utility indicates one bundle is preferred to the other. The consumer is indifferent between two bundles if each gives the same level of utility.

The other simplification is to assume that there is only one "representative" consumer in the economy. Using this short-cut allows analysis in terms of only one consumer, because all consumers are identical. It is easy to criticize this assumption, but it can be justified if one is less interested in the differences between consumers in an economy and more interested in other features. For example, international macroeconomic models are often built to analyze differences between countries, not within each country. Simplifying to a representative agent allows use of the current modeling structure.

The Budget Constraint

Using these simplifying assumptions, we can begin with characterizing what is feasible for the representative consumer. In the current framework, this amounts to specifying the different constraints of the consumer. In outlining these constraints, we assume that the consumer is a price taker in goods and labor markets. All such prices are specified in real terms, specifically in terms of the consumption goods. This is because there is no money in the current model. Introducing money into the general equilibrium framework can be done, but raises some methodological issues.

The consumer's first constraint, the time constraint, says that the time devoted to work plus leisure

time must equal the total time available:

$$l + N^s = h \tag{1}$$

This constraint specifies N^s as labor supply (or the time available for work) and h as total time available. The constraint is kept general so that the total time available can be changed according to the time period under consideration (i.e. daily, weekly, etc.). The second constraint is a budget constraint, which states that income equals expenditures. Income consists of labor earnings plus dividends from the ownership of firms less any lump-sum taxes. Expenditures are only for current consumption goods:

$$p_c C = wN^s + \pi - T \tag{2}$$

The left-hand side is total consumption expenditures on consumption goods, where p_c is the price of such goods. It is common practice to normalize this price to 1, and divide all other prices by this value. This makes the price of consumption goods the numeraire, meaning that all other goods are specified in terms of consumption goods. For example, the wage rate (w) is not dollars per hour, but consumption goods per hour when p_c is the numeraire. Total income depends on labor income, wN, where N is labor supply. The consumer is also assumed to own any firms in the model, so collects any dividends π as income. Taxes are collected by the government, but only in a lump-sum manner (T) whereby they do not change any of the consumer's decisions on the margin.

The two constraints can be combined into one general budget constraint:

$$C = -wl + wh + \pi - T \tag{3}$$

Notice that p_c and labor supply are no longer in the equation. This overall budget constraint states that consumption in the current period is limited to disposable income, which is the sum of labor and dividend income less taxes. The power of rewriting the budget constraint in this manner is that it can graphically depict the relationship between consumption of physical goods and leisure, the two goods which comprise the consumption bundle. Figure 1 plots this with C on the vertical axis, l on the horizontal, and a slope $-w$.

The shaded area below the budget constraint constitutes feasible consumption bundles. If the consumer chooses not to take any leisure (l=0), then the amount $wh + \pi - T$ can be consumed, which is the vertical intercept. If the consumer chooses to take the maximum amount of leisure (C=0), then the horizontal intercept is $h + \frac{\pi - T}{w}$. If $\pi - T < 0$, then the horizontal intercept will be less than h, as shown in Figure 1. In the case where $\pi - T > 0$, the horizontal intercept will be greater

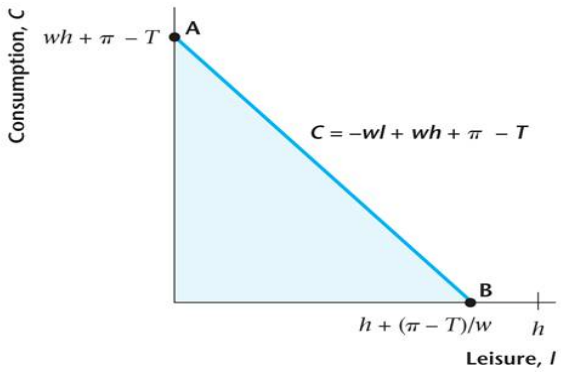


Figure 1: Representative Consumer's Budget Constraint with $T > \pi$, from Williamson (2011)

than h, which is not possible because h is the maximum time available. In this case the budget constraint will be kinked at h as shown in Figure 2.

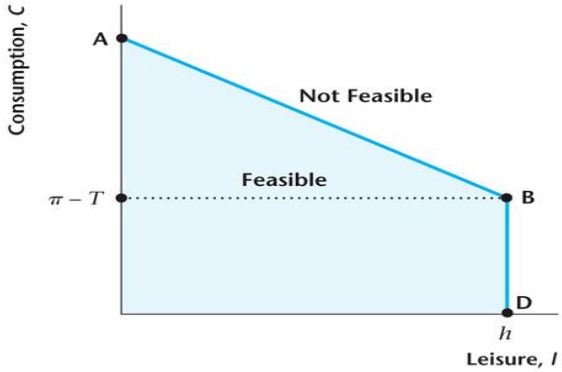


Figure 2: Representative Consumer's Budget Constraint with $T < \pi$, from Williamson (2011)

In general it makes no difference which case is used, but the kinked budget constraint will be considered the base case. Figure 2 also shows that if all time is spent as leisure, then the consumer still has consumption equal to $\pi - T$, due to dividend income from firms because profits exceed taxes. As a final point, consider the slope of the budget constraint, $-w$. As mentioned earlier, this is in terms of the physical consumption good, so can be interpreted as consumption goods per unit time of labor supply. In this sense it is also the tradeoff between consumption and leisure. An extra unit of leisure results in a loss of w goods in income, and similarly a reduction of a unit in leisure results in an extra w goods in income. Another way to say this is that the wage rate is the opportunity cost of leisure, a point that will be important when the wage rate changes.

Preferences

The utility function of the representative consumer depends on consumption of a physical consumption good and leisure. Before this function can be used in our analysis, there are some additional assumptions on its properties which must be made. The assumptions are pragmatic because they are necessary to find a unique solution to the consumer's optimization problem, but do have some economic justification.

The first assumption is that more is preferred to less. That is, additional amounts of the physical good or leisure give a higher level of utility. This is a reasonable simplification, although it won't always be true. Consumers are also assumed to prefer diversity in their consumption bundle, also called the property of convexity. If the consumer is indifferent between two consumption bundles, then some mixture of the two bundles is preferable to either. Again, this is reasonable but won't always be true. An implication of this is that there is decreasing marginal utility of either the physical good or leisure. Each additional unit of either the physical good or leisure consumed raises utility, but raises it by smaller and smaller amounts. The final assumption is that consumption and leisure are both normal goods. This means that consumption and leisure increase and decrease with income.

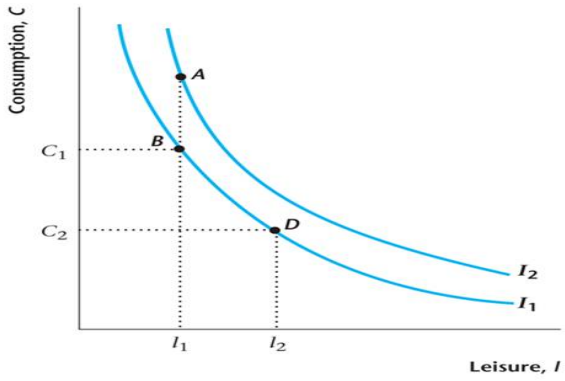


Figure 3: Indifference Curves, from Williamson (2011)

Because of these assumptions, the static nature of the model, and the fact that there are only two goods, the consumer's preferences can be summarized using indifference curves as shown in Figure 3. These curves are plotted with consumption on the vertical axis and leisure on the horizontal. The level of utility along each indifference curve is the same, and each curve is bowed-in towards the origin (convex). Each curve represents all of the bundles of consumption and leisure which give the same level of utility.

The fact that consumers prefer more to less means that the level of utility rises as one moves north-east in the diagram. To see this, take point B in the diagram and compare it with point A. Each bundle has the same level of leisure (l_1), but point A has a higher amount of the consumption good. By assumption more is preferred to less, so point A must yield a higher level of utility than

point B. And because each indifference curve plots bundles with the same level of utility, all of the bundles that comprise I_2 must have a higher level of utility than those which comprise I_1.

The preference for diversity is what gives each indifference curve its bowed-in shape. This is best understood by introducing the marginal rate of substitution of leisure for consumption (MRS_{lc}), which is the slope of the indifference curve. Its value gives the rate at which the consumer is willing to substitute leisure for consumption goods. Begin near the top of I_1 at point B. At this point the slope is relatively steeper, or the absolute value of the MRS_{lc} is higher than at other points on the curve. The interpretation is that the consumer is very willing to substitute consumption goods for leisure. To see this, beginning at point B, move to the right by one unit of leisure. To keep the same utility (i.e. to stay on the same indifference curve), the consumer would be willing to give up dC_1 units of consumption. It seems that the consumer is willing to give up a substantial amount of C to get an additional unit of leisure in this case, thus the steeper slope.

Contrast this by starting at point D in Figure 3, adding one unit of leisure, and then calculating dC_2. Notice that dC_2 is substantially smaller than dC_1, which gives a lower MRS_{lc}, and means that the consumer is less willing to give up consumption for leisure in this case. Thus the marginal rate of substitution is decreasing as one moves down the indifference curve, which gives the curves their shape. This implies a preference for diversity because when consumers have more of the consumption good (as at point B), they are more willing to trade that good (higher MRS_{lc}) than when they have relatively less of it (lower MRS_{lc}, as at point D).

The Optimal Bundle

At this point the bundles which are feasible for the consumer and the consumer's preferences over those bundles have been characterized. The next task is to determine how the consumer chooses the most preferred bundle from what is available. This is often called the consumer's optimization or maximization problem. In this case the solution can be shown graphically, as in Figure 4.

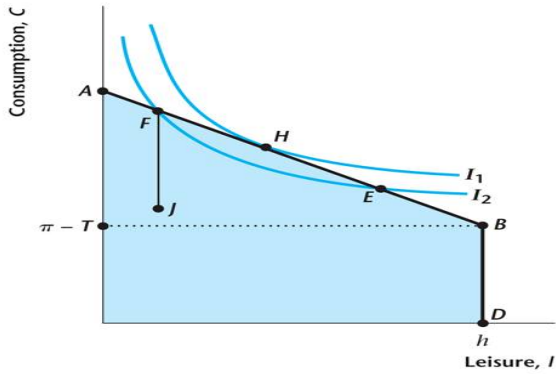


Figure 4: The Optimal Bundle, from Williamson (2011)

Notice that this diagram combines the two diagrams of the consumer's budget constraint and indifference curves shown above. The optimal bundle is at point H in the figure. Why is this optimal? Recall that utility increases as we move north-east in the diagram, which means that the consumer would like to choose the highest indifference curve possible. However, the feasible bundles are bounded by the consumer's budget constraint. Thus, the optimal bundle will be a point on the budget constraint (this is the most one can consume given income) that just touches an indifference curve.

The optimal bundle must also have the property that the slope of the budget constraint is equal to the slope of the indifference curve at the point where they touch. On the figure we would say that the indifference curve is tangent to the budget constraint. If true, this means that $MRS_{lc} = w$, or the rate at which the consumer is willing to trade consumption for leisure is the same rate at which the market is willing to trade consumption for leisure. Recall that the wage rate is the opportunity cost of leisure, also the value of leisure in terms of consumption. And the marginal rate of substitution of leisure for consumption is the rate at which the consumer is willing to trade consumption for leisure.

Figure 4 shows two other points where the indifference curve touches the budget constraint which are not tangent. At point F, the slope of the indifference curve is higher than the slope of the budget constraint. That is, the consumer is more willing to trade consumption for leisure than the market, or the $MRS_{lc} > w$. This cannot be an optimal solution because the consumer can move to a higher indifference curve through trade. This is done by giving up w units of consumption (the market currently values consumption more than the consumer) for 1 unit of leisure. This is a good trade for the consumer because they would have been willing to give up $MRS_{lc} > w$ units of consumption goods instead. By continually making such trades the consumer will move along the budget constraint until point H, where the wage rate and marginal rate of substitution are equal.

The same is true beginning at point E. In this case, $MRS_{lc} < w$, or the consumer is less willing to trade consumption for leisure than the market. As before, the consumer can move to a higher indifference curve through trade. This is done by giving up one unit of leisure in exchange for w units of consumption goods. The consumer will make this trade because they would have been willing to take $MRS_{lc} < w$ units of consumption goods for that unit of leisure. The consumer can then move along the budget constraint towards a higher indifference curve until the marginal rate of substitution equals the wage rate.

From either direction the consumer moves towards the point where the indifference curve is tangent to the budget constraint, which is the optimal bundle. If the consumer is not at this point they can always improve their situation through trade, eventually moving to a higher indifference curve, and towards this optimal bundle. Keep in mind that this nice neat solution only applies because of the assumptions made on the preferences above.

Income and Substitution Effects

Once the optimal bundle has been chosen, our attention turns to what happens when the consumer's income changes or market prices change. This can be handled by breaking down any changes into income and substitution effects. Income effects are exactly that, the optimal responses of the consumer when income varies. Substitution effects result from the changes in the relative price of one good versus another, which can lead to substitution between them.

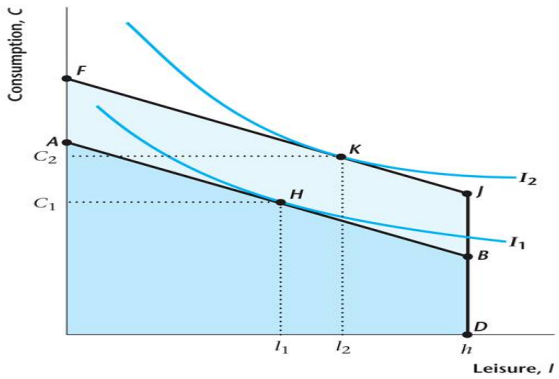


Figure 5: A Pure Income Effect, from Williamson (2011)

Begin with an increase in dividend income or a decrease in taxes for the consumer so that $\pi - T$ rises. Assume that the wage rate and price of consumption goods does not change. This is an example of a pure income effect. The first implication of this change is that the consumer's budget constraint will shift out. This is because the higher income makes additional consumption bundles feasible for the consumer. Mechanically, the rise in $\pi - T$ shifts up the vertical intercept ($wh + \pi - T$), and the entire constraint must shift up because the wage rate stays the same.

Figure 5 shows this shift upwards, and also the corresponding change in consumption bundles to the one at point K. A pure income effect allows the consumer to increase consumption of either the physical good or labor, which results in higher utility. Notice also that point K is both higher than and to the right of the original bundle, point H. This is because we assumed that both consumption goods and leisure are normal goods, so that consumption of both must increase with income. Choosing the new bundle at point K reflects this fact.

The result is that higher non-wage disposable income increases consumption and decreases labor supply. Again, this is reasonable but won't always be true. In Figure 5 this is reflected by the fact that the increase in income is the difference between points A and F, but the increase in consumption is $C_2 - C_1$, which is less than the increase in income. This occurs because the higher disposable income has reduced labor income as leisure has increased, which is reflected by the fact that consumption does not increase one for one with income.

Things become more complicated when the wage rate changes, holding everything else constant. The

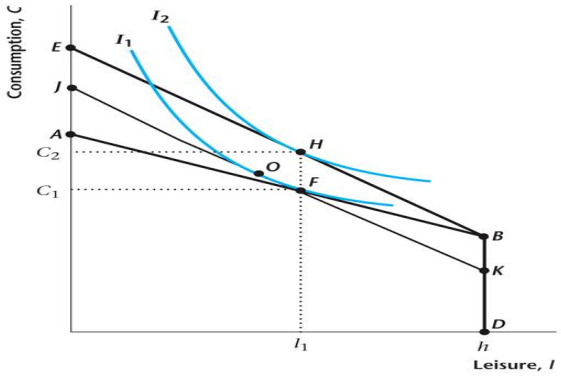


Figure 6: Income and Substitution Effects, from Williamson (2011)

implications of this change can be divided into a substitution effect and a pure income effect. In Figure 6, begin with the budget constraint ABD and assume that the wage rate rises. This higher wage rate increases the slope of this constraint to something like EBD. The original optimal bundle was F and has moved to something like H. The key point to illustrate here is that with a rise in the wage rate consumption must increase, but leisure can increase or decrease depending on magnitude. This differs from the pure income effect where both consumption and leisure rose.

This occurs because the rise in the wage rate creates a tradeoff. Because the wage rate represents the opportunity cost of leisure in terms of consumption goods, a rise in its value means that leisure is more expensive relative to consumption. Consumers can be expected to substitute away from leisure (to work more) towards consumption. Think of the wage rate as the price of leisure, which has just risen. In the diagram this is depicted by rotating the curve ABD to JKD, which has the new, higher wage rate as its slope. Intuitively, this is like asking how much leisure can I take away and keep the same level of utility at the new wage rate. The point O is the new optimal bundle reflecting only the substitution effect. As expected, consumption has risen and leisure has fallen.

The rise in the wage rate also leads to a pure income effect. This is because a higher wage rate increases the consumer's income, all else equal. Graphically, this results in a shifting of the budget constraint from JKD to EBD. Because this is a pure income effect the consumer will increase consumption of both the physical good and leisure. Another way to think about this effect is that the increased wage rate means the consumer can work less and still get the same amount of income, which would lead the consumer to increase leisure. The sum total of both substitution and income effects means that the consumer will increase consumption of physical goods, but it is uncertain what happens to leisure, and correspondingly labor supply in the face of an increase in the wage rate.

The Labor Supply Curve

The income and substitution effects described above can be used to derive a relationship between the wage rate and employment, or the labor supply curve. This is the amount of labor the consumer wishes to supply at any given wage rate. It can be derived by considering the time constraint, slightly rearranged:

$$N^s(w) = h - l(w) \tag{4}$$

In this case the dependence of labor supply and leisure on the wage rate has been made explicit. This also provides the mechanism whereby a relationship between the wage rate and labor supply can be established. One could begin at some wage rate w^* and then increase the wage rate to w^{**}. How does labor supply change? From above, we know that the impact of an increase in the real wage rate is ambiguous on the demand for leisure. However, in what follows we assume that the substitution effect dominates the income effect. This means that a higher wage rate will decrease leisure, or increase labor supply, giving the upward-sloping labor supply curve in Figure 7.

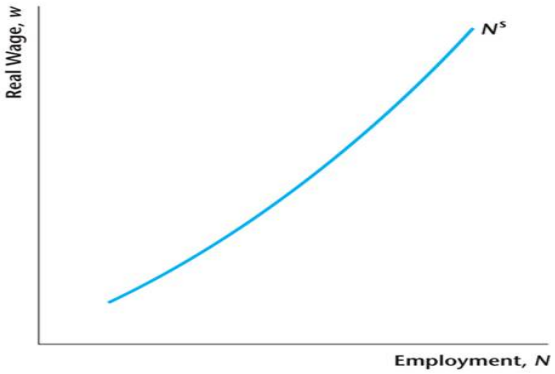


Figure 7: The Labor Supply Curve, from Williamson (2011)

This assumption is intuitive, and gives the usual upward sloping labor supply curve, but depends on the strength of the income effect. This is an active area of research, and may vary between countries. The slope of this curve is given by the strength of the substitution effect. The steeper the slope, the less responsive is the consumer to changes in the wage rate, the flatter the slope, the more responsive is labor supply to changes in the wage rate.

Also, the location of the labor supply curve depends on the consumer's income. For example, an increase in income results in both higher consumption and higher leisure. This higher leisure reduces labor supply at a given wage rate, which shifts the curve to the left as shown in Figure 8.

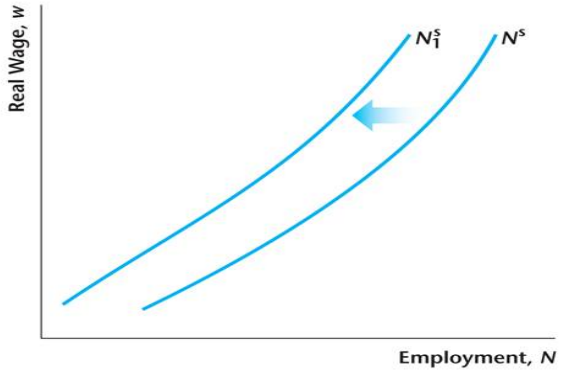


Figure 8: A Shift in the Labor Supply Curve, from Williamson (2011)

Firms

The characterization of firms in general equilibrium models is based on a particular theory of the firm. This neoclassical theory views firms as perfectly competitive entities which make decisions by either maximizing profits or minimizing costs (they are equivalent under certain conditions). This optimization problem is subject to various constraints, the most important of which is the production technology available to the firm.

As with the consumer, we simplify the firm's problem by assuming a representative firm. Using this short-cut allows analysis in terms of only one firm, because all firms are identical. It is easy to criticize this assumption, but it can be justified if one is less interested in the differences between firms in an economy and more interested in other features, as with consumers. The other gross simplification is that this one firm produces only one consumption good, which is what is available to the consumer. This limits the current model to one sector, making graphical analysis possible. Either of these restrictions can be lifted, but at the cost of substantial complication which is not discussed here.

The Production Function

Before outlining the full profit maximization problem of the firm, it is important to fully understand the firm's production function. This is assumed to depend only on capital and labor demand, and can generally be written:

$$Y = ZF(K, N^d) \tag{5}$$

In this equation, Y is the output of consumption goods, z is total factor productivity or technology, K is the quantity of capital input to the production process, and N^d is the quantity of labor input measured as total hours worked by employees of the firm. In other words, production is a function of capital and labor.

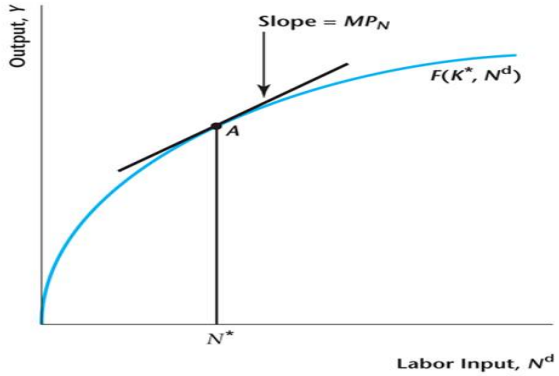


Figure 9: The Production Function, from Williamson (2011)

In the current static set-up we assume that the capital input is fixed, so only labor can be varied. The total factor productivity (TFP) captures the degree of sophistication of the production process. An increase in Z makes both factors more productive, while a decrease has the reverse effect. This is often referred to as technology, but can include other intangible things like the weather, management skill, etc.

As with the utility function, there are several assumptions placed on the production function. Each is reasonable, but may not hold in all cases. The first is that the production function exhibits constant returns to scale (CRS). This means that if all factor inputs (in this case K and N^d) are changed by some factor λ, then output changes by the same factor. Mechanically:

$$ZF(\lambda K, \lambda N^d) = \lambda Z F(K, N^d) \tag{6}$$

This assumption means that small firms and large firms are equally as efficient in producing output from factors of production. Increasing returns to scale, where changing the inputs by λ results in an output increase greater than λ, implies that large firms are more efficient than small firms. This is because greater inputs lead to higher and higher output. The reverse is true with decreasing returns to scale. Empirically, it is unclear whether the economy as a whole is characterized by constant returns to scale. However, this assumption allows us to use the representative firm, because a small firm is the same as a very large firm, which is the representative one.

The second assumption is that output is increasing in either of its two arguments. Increasing the amount of capital or labor will raise the amount which can be produced. Another way to say this is that the marginal product of either factor, the increased production from a one unit increase in either factor, is positive for both inputs. This is highlighted in Figure 9.

There are two aspects of Figure 9 which stand out. The first is that there is a positive marginal product of labor, which is also the slope of the production function. The second is that the shape is

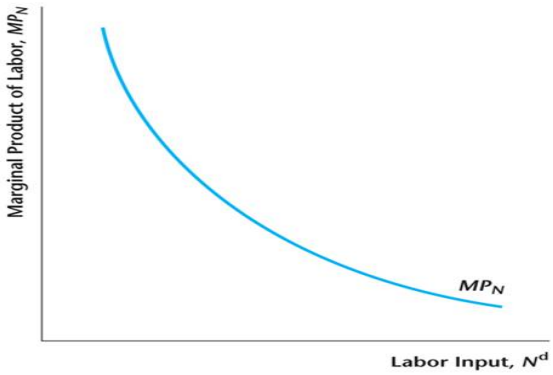


Figure 10: The Marginal Product of Labor, from Williamson (2011)

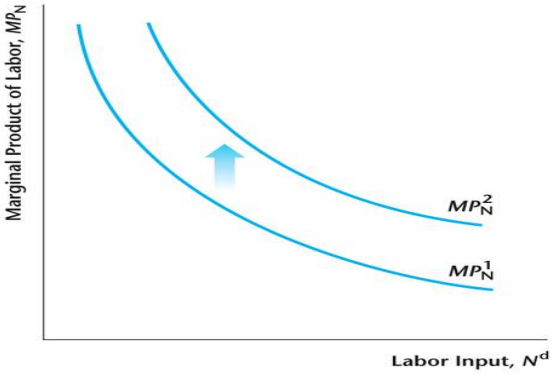


Figure 11: A Shift in the Marginal Product of Labor, from Williamson (2011)

bowed, or concave. This second feature is the next assumption made on the production technology, that the marginal product itself decreases as the amount of labor supplied increases. That is, adding an extra unit of labor increases production, but at lower and lower rates as more labor is added. Figure 10 plots the marginal product of labor against the labor input.

As final assumption is that the marginal product of labor increases when the level of capital increases, all else equal. One can think of this as each worker having access to more machines in order to get their work done. This will tend to make them more productive, and results in a shift in the marginal product curve upwards as in Figure 11.

These last two properties also apply to the marginal product of capital as well. With these assumptions and the graphical representation, we can asses the impact of a change in TFP on output and the marginal product of labor. If total factor productivity rises, say due to technological progress, we would expect more output to be produced given the same inputs of capital and labor. Diagrammatically, as shown in Figure 12, the level of output for any given level of labor input is now higher.

A good way to think about this is that the production possibilities of the firm are now greater because

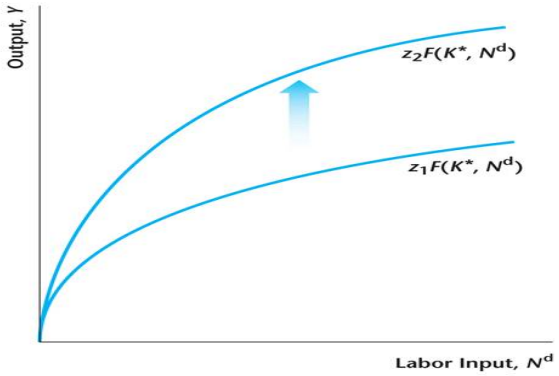


Figure 12: An Increase in TFP, from Williamson (2011)

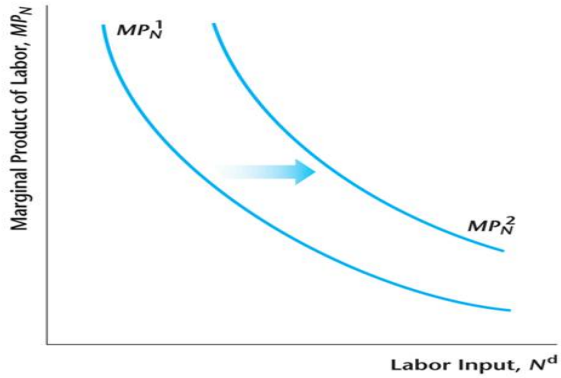


Figure 13: An Increase in TFP Cont., from Williamson (2011)

of the technological growth. Notice that the rise in TFP is not symmetric on the production function, as the slope at every N^d has risen. This makes sense because greater TFP allows for higher output with the same level of inputs, or a higher marginal product of labor. The result is that the marginal product of labor curve shifts as well, as shown in Figure 13.

Figure 13 can be interpreted as showing that the marginal product of labor is higher at any level of labor demand due to the TFP growth.

Profit Maximization

With the production technology and its properties in hand, the profit maximization problem of the firm can now be outlined. In the static case, with capital stock fixed, this amounts to choosing the amount of labor to demand to maximize the difference between total revenues and total costs. The revenues of the firm are the value of its sales, the number of consumption goods sold times their price. The costs in this case are only the wage rate paid for labor, as capital costs are assumed to be paid in the past.

This procedure is depicted graphically in Figure 14. The line beginning at point D represents the

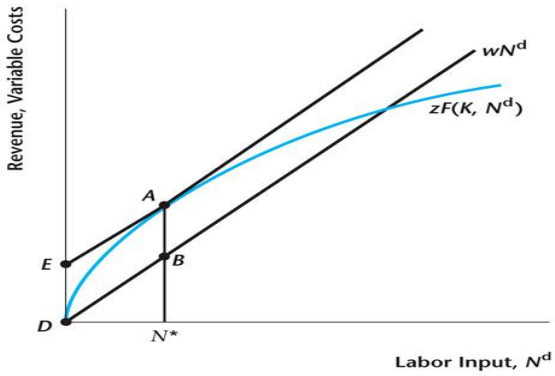


Figure 14: Profit Maximization, from Williamson (2011)

firm's costs, which are the wage rate times hours worked. As shown, these increase at a constant rate of w as each unit of labor is added. The revenue of the firm is its production (also equal to its sales) times price, which is 1 by assumption because it is the numeraire. Thus we can use the production function to show the revenues of the firm. Our goal then it to maximize the difference between these two lines, which occurs at N^*, and is the distance AB. It turns out this is the point where the slope of the production function, the marginal product of labor, equals the slope of the costs, the wage rate. That is, where the wage rate equals the marginal product of labor.

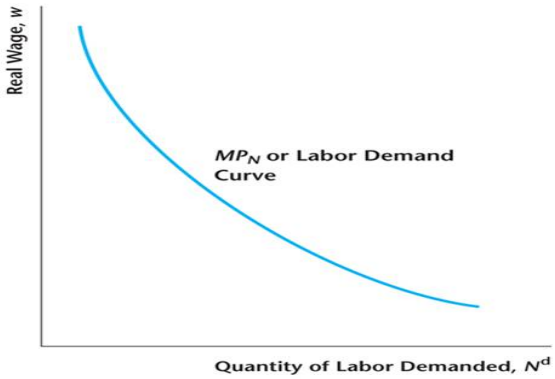


Figure 15: The Labor Demand Curve, from Williamson (2011)

The intuition for this result is very clear. If the marginal product of labor is higher than the wage rate, this means that the benefit to firms (the marginal product) outweighs the costs (the wage rate). It makes sense in this case to hire more labor, which works to reduce the marginal product. Alternatively, if the marginal product of labor is lower than the wage rate, the costs outweigh the benefits, and it makes sense for the firm to reduce workers, increasing the marginal product. The optimal point for the firm is where the benefit of a worker just equals the costs, or where $MP_N = w$, which is N^* in Figure 14.

This result also allows us to derive the labor demand curve of the firm. Because the marginal product equals the wage, the plot of the marginal product against labor demand is the labor demand curve. This gives the relationship between the wage rate (which must equal the marginal product of labor for the profit maximizing firm at an optimum) and the amount of labor demanded, as shown in Figure 15.

Any profit earned by the firm is represented as π in the consumer's budget constraint, as the consumers own the firms.

Government

The representation of government in the static model is basic. The government derives its revenue by levying a lump-sum tax on consumers, T. This revenue is used to finance purchases of consumption goods, G. Here, we let G be determined exogenously, or outside the model. Another assumption made on government is that it has a balanced budget, or that its expenditures must equal its revenues:

$$G = T \tag{7}$$

This is also known as the government budget constraint.

Equilibrium

The individual decisions of consumer and firms, along with the exogenous government, are brought together using the concept of general equilibrium. When the representative firm is assumed to be perfectly competitive, this is also called a competitive equilibrium. Once this has been done, the full model is depicted graphically in the static case, and some experiments using the model are conducted.

Defining a Competitive Equilibrium

One could easily imagine a situation where the amount of goods produced by the firm is different than the amount demanded by consumers, or where the labor supply of consumers varies from labor demanded by firms. Defining an equilibrium in a clear manner avoids this situation.

Definition. *A general equilibrium in the current model is the values of endogenous quantities (C, N^s, N^d, T, and Y) and an endogenous price (w), given exogenous quantities (G, Z, and K) such that the following conditions are satisfied:*

1. *The consumer optimizes given market prices;*
2. *The firm optimizes given market prices;*
3. *The labor market clears, $N^s = N^d$; and*
4. *The government budget constraint is satisfied, $G = T$.*

Notice the detail embodied in this definition. An equilibrium consists of the values of the endogenous variables, both prices and quantities. These are the items of interest, and the exogenous quantities are taken as given. However, these solutions are only valid if the four specified conditions hold. The first two state that the solution must be optimal for both the consumer and the firm. The consumer is choosing a bundle where their indifference curve is tangent to their budget constraint. And the firm is choosing to produce where the difference between total revenue and total cost is greatest.

The third condition ensures that the labor market clears. So the quantities and prices which comprise an equilibrium must be consistent with optimization by both consumers and firms, and they also must be specified so that labor supply and demand equate. The final constraint forces the government to meet their constraint as well. There is a fifth constraint which is implied by the first four, namely that the consumption goods market clears, $Y = C + G$. To see that this is implied, begin with the consumer's budget constraint:

$$C = wN^s + \pi - T \tag{8}$$

Because the firm optimizes, $\pi = Y - wN^s$, and by assumption $G = T$ so this becomes:

$$C = wN^s + Y - wN^d - G \tag{9}$$

The labor market must clear in a competitive equilibrium, so $N^s = N^d$, which leaves the implied goods market clearing condition:

$$C = Y - G \tag{10}$$

It is important to emphasize that defining a general equilibrium is what gives the model consistency. Without the specific definition, there is no means of verifying that the quantities demanded/supplied will be the same between the consumer and firm.

Representing the Equilibrium Graphically

The power of beginning with a static model which has a representative consumer and a representative firm is that the model can be displayed graphically. When exogenous variables are changed, these graphs can also be used to see how the endogenous variables change. We have already derived the graphical representation of the consumer's problem, the indifference curves and budget constraints, but need to do so for the firm as well using the same axis (C, l) as that of the consumer.

To do this, begin with the firm's production technology, $Y = ZF(K, N)$. Notice that because we have assumed an equilibrium, N^d is now $N = N^s = N^d$. Using the time constraint, the amount of labor

can be substituted out of the production function to give $Y = ZF(K, h - l)$.

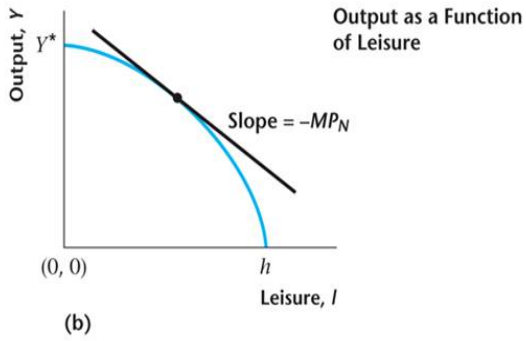


Figure 16: A Different View of the Production Function, from Williamson (2011)

This gives a production relationship between output and leisure. Graphically, it amounts to flipping around the production function, as shown in Figure 16. In this case, if leisure is on the horizontal, and $l = h$ or $N = 0$, there will be no output produced. Similarly, if $l = 0$ or $N = h$, then the maximum output will be produced. In this way the production function shown in Figure 16 is derived. The final step is to change the vertical axis from Y to C to make it consistent with the axis used for consumption.

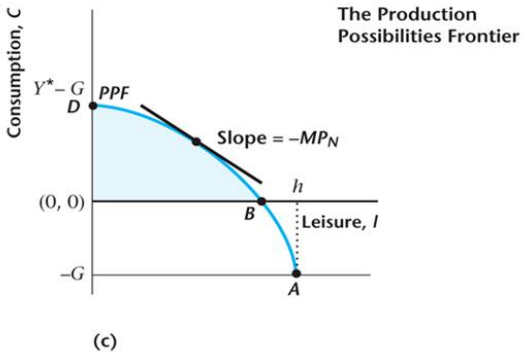


Figure 17: The Production Possibilities Frontier, from Williamson (2011)

This can be done by using the market clearing condition in the goods market, $Y = C + G$ or $C = Y - G$. Subtracting G from every point on the reversed production function gives a curve which can directly be compared with the consumer's indifference curves and budget constraint, and this is shown in Figure 17. This modified curve is also known as the production possibilities frontier (PPF).

PPFs in general show the production tradeoff between two commodities in an economy given fixed factors of production. The PPF in this case depicts the tradeoff between consumption and leisure with fixed capital, but where labor supply/demand vary. We know that consumption cannot be negative so

any combinations to the right of point B are not feasible. Also notice that the slope of the PPF is the negative of the marginal product of labor, or the increment to output with an addition of one unit of labor. This value is also called the marginal rate of transformation (MRTS). It represents the rate at which leisure can be converted to consumption goods using available production technology, and is denoted $MRTS_{lc}$ in this case.

The consumer's indifference curves and budget constraint can now be put together with the firm's production possibilities frontier to show the general equilibrium solution, as in Figure 18.

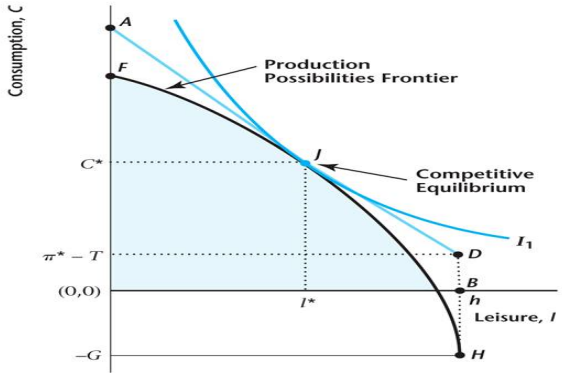
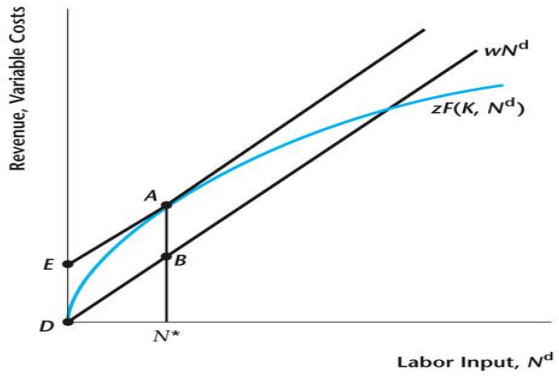


Figure 18: The Competitive Equilibrium, from Williamson (2011)

Why is point J the optimal solution? Look back at the conditions which constitute a competitive equilibrium. First, the consumer must be optimizing. At point J, the $MRS_{lc} = w$, which is the condition for consumer optimality outlined above. The firm must also be maximizing profit at any competitive equilibrium. At point J, the $MP_N = w$, which is the condition for firm optimality outlined above. Meeting these first two conditions implies that $MRS_{lc} = w = MP_N = MRTS_{lc}$. The marginal rate of substitution of leisure for consumption and the marginal rate of transformation of leisure for consumption must be equal. This makes sense given what each of those rates represent, as this is where the consumer's value of the leisure/consumption tradeoff (w) is the same as the firm's, which gives neither a reason to deviate from this point.

Point J also meets the third condition that labor supply equals labor demand, as both labor supply and demand are $h - l^*$, as shown in Figure 18. Finally, the government also meets it budget constraint in this case. This last condition is more difficult to show. In order to do so, the first step is to derive the profit given labor input of l^*, which is the distance DH in Figure 18. We know this is the profit from Figure 14 (shown here again for ease of comparison).

The optimal solution occurs where profit is equal to AB, but this is also the same as ED in the figure. Once the production function is flipped, along with the line which is tangent to point A, we can see that the distance ED is actually the same as the distance DH in Figure 18. Once this has been established, we know the distance BH must be equal to $-G$, so that DB is equal to $\pi - G$. But



Profit Maximization, from Williamson (2011)

point D also represents the consumer's dividends less taxes, as AD is the budget constraint. Taken together these imply that $G = T$. This verifies that the figure depicts a competitive equilibrium, and the model can be used to conduct hypothetical scenarios.

Example: A Change in G

The first experiment we conduct is to increase government spending in the model. Increasing government spending from G_1 to G_2 immediately shifts down the PPF. This is because $C = Y - G$ in the model. This reduction in the production possibilities for the firm is a negative income effect, as any increases in G must be offset by higher taxes because $G = T$ in equilibrium. The negative income effect results in less demand for both consumption and leisure.

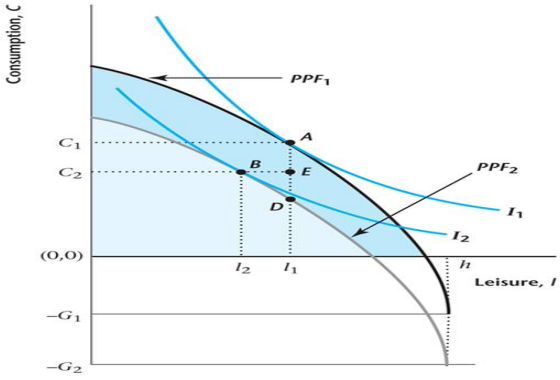


Figure 19: An Increase in Government Spending, from Williamson (2011)

This is depicted in Figure 19, where the initial point is at A and the final point after a shift down in the PPF is at B. Notice that consumption falls by the distance AE, which is less than the rise in G, depicted by AD. This occurs because the negative income effect leads to additional labor supply, which means that output must rise. This is an important point. Output rises with an increase in G in

the model because labor input rises.

While the slope of the PPF does not change, this is not true of the real wage. We know the real wage must fall because labor supply rises, which only happens with a lower wage. This can also be deduced from Figure 19. Because the new equilibrium point is to the left of the original one, and we know the marginal product of labor falls with more labor input, the real wage must fall. Recall that moving to the left in this figure is lower leisure, or more labor.

In summary, a rise in government spending reduces consumption and leisure, but increases output and hours.

Example: A Change in Z

The second experiment, a rise in TFP, is slightly more complicated. A rise in TFP will shift up the production function, which is equivalent to shifting out the PPF. As was mentioned above, this shift also results in a higher marginal product at each level of labor input, so the slope of the PPF at each l also changes. This change in the slope means that the wage rate facing both consumers and firms must change, which induces both income and substitution effects in response.

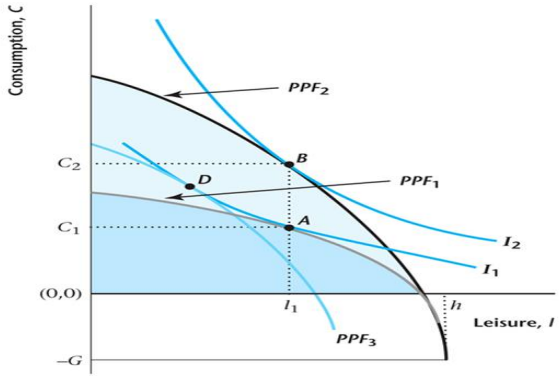


Figure 20: An Increase in TFP, from Williamson (2011)

Figure 20 shows these effects, along with the final result. The immediate increase in z will raise w. Because the wage rate represents the opportunity cost of leisure, the consumer will substitute away from leisure towards consumption. This substitution effect is shown in the figure as PPF_3, which is a rotation of the original PPF. Point D shows the bundles of C and l which keep the original level of utility but reflect the new wage rate.

There is also an income effect, as the consumer now has greater income if they choose to work the same amount. Because consumption and leisure are normal goods, demand for each will increase. This is shown as a shifting out of the PPF to PPF_2. Point B is not necessarily the solution, it shows that C must rise, but l is ambiguous due to the offsetting income and substitution effects. The real wage rises in this experiment as well. This is true because point D in the figure has a higher slope than

point A, as the MRS_{lc} rises when moving up an indifference curve. But any point to the right of D on the PPF must be steeper still, as the marginal product of labor is rising as l rises. When PPF_3 shifts up to PPF_2, the slope remain the same, so this must still hold, giving a higher real wage rate.

Extension to a Small CGE Model

The small model outlined to this point is very useful for introducing general equilibrium and conceptualizing various experiments, but does not contain sufficient detail for many purposes. Computable general equilibrium (CGE) models are extensions of this basic model to incorporate much more detail. With such models it is no longer possible to depict a general equilibrium graphically, so numerical simulation are used instead. This section introduces two small CGE models, to illustrate their structure. The first model represents a closed economy with two sectors, two factors of production, and one representative agent. The second model extends this framework to an open economy by adding another identically structured region.

Closed Economy

The model is of a closed economy and has one representative consumer. Unlike the model above, there are two sectors, each with a representative firm, which means there are two goods which are produced. There is no leisure in this model. The competitive equilibrium concept is the same as that described earlier.

The consumer problem is now to choose each good which is produced (C_x and C_y) to maximize utility. There are two equivalent ways to proceed. One way is to define the consumer's utility over both consumption goods, $U = U(C_x, C_y)$. The other is to construct a composite consumption good, $\hat{C}$, and then define utility over this composite, $U = U(\hat{C})$. As more goods are added to a model the composite construction is easier to deal with, and that is the approach used below.

Simulating CGE models is a quantitative exercise, so an explicit functional form for utility is needed for the consumer's optimization problem. The simplest form to use is logarithmic utility, $\log \hat{C}$. The consumer's problem then reads:

$$\begin{aligned} &\max_{\hat{C}} \quad \ln \hat{C} \\ s.t. \quad &w(N_x + N_y) + r(K_x + K_y) = M = \hat{p}\hat{C} \end{aligned} \tag{11}$$

The consumer optimizes by choosing consumption of the composite good subject to their budget constraint. The budget constraint equates income (M) to expenditures. Expenditures result from purchasing the composite good at price $\hat{p}$. Consumer income is derived from labor, wN where w is the wage rate and N is labor supply to each sector, and capital income, rK where r is the rate of return on capital and K is the capital stock in each sector. This set-up assumes that the consumer owns the

capital stock and rents it to firms at price r per unit.

There is a second step to this optimization because the composite consumption good is a modeling construct. The consumer must minimize total costs of consumption by choosing the amount of good X (C_x) and good Y (C_y) to consume. This is subject to an "aggregator" of the two goods, where p_x and p_y are the prices of good X and good Y:

$$\begin{aligned} &\min_{C_x, C_y} \quad p_x C_x + p_y C_y \\ s.t. \quad &\hat{C} = [\gamma_c C_x^{\rho_c} + (1-\gamma_c) C_y^{\rho_c}]^{\rho_c} \end{aligned} \tag{12}$$

The constraint in this equation is a constant elasticity of substitution (CES) function. These types of functions are common in general equilibrium models. This function provides the means by which the individual goods can be combined to form a composite. The γ_c are share parameters, which dictate how much of each good is consumed in this aggregation. The ρ_c is a parameter which represents the elasticity of substitution between the two types of goods. Specifically, $\rho_c = \frac{(\sigma_c - 1)}{\sigma_c}$ where σ_c is the elasticity of substitution between C_x and C_y.

The elasticity of substitution is a key parameter in general equilibrium modeling. In this case, it quantifies the willingness of the representative consumer is to substitute between the two goods in consumption when prices change. A higher value means that the consumer is more willing to substitute. Technically, this elasticity is the curvature of the consumer's indifference curve (or isoquant when used as a production function). The CES function allows the modeler to choose a value for σ_c, and this value stays fixed throughout the simulation.

The Cobb-Douglas form is an alternative which could be used in lieu of the CES.[3] This has the form $\hat{C} = C_x^{\alpha} C_y^{(1-\alpha)}$ in this case, where α is a share parameter. Using Cobb-Douglas is simpler mathematically. However, the implied elasticity of substitution between consumption goods based on the Cobb-Douglas form is one, which is not ideal in many cases.

Solving both the first and second stages of the consumer's problem (with a CES aggregator) leads to three optimality conditions which must be met:

$$C_x = \gamma_c \hat{C} \left(\frac{\hat{p}}{p_x} \right)^{\sigma_c} \tag{13}$$

$$C_y = (1-\gamma_c) \hat{C} \left(\frac{\hat{p}}{p_y} \right)^{\sigma_c} \tag{14}$$

[3] Technically, the Cobb-Douglas form is a special case of the CES when the elasticity of substitution is one.

$$\hat{p} = [\gamma_c p_x^{(1-\sigma_c)} + (1-\gamma_c) p_y^{(1-\sigma_c)}]^{\frac{1}{(1-\sigma_c)}} \tag{15}$$

These are the conditions which must hold for there to be any competitive equilibrium. The first two specify that demand of either good is a fraction of total consumption based on relative prices, consumption shares, and willingness to substitute. The third shows that the composite price depends on these same factors.

Each firm's problem is comparatively simple, it chooses capital and labor to maximize profit (the industry X is shown here):

$$\begin{aligned} \max_{K_x, N_x} \quad & p_x Q_x - w N_x - r K_x \\ s.t. & Q_x = Z_x K_x^{\alpha_x} N_x^{(1-\alpha_x)} \end{aligned} \tag{16}$$

In this problem, Q_x is production of good X, Z_x is total factor productivity in the production of good X, and α_x represents the capital share of output. Notice that the production function in this case has a Cobb-Douglas form, which is standard, although CES could also be used. There are no intermediate goods here, but these could also be introduced, as could a multi-stage optimization as with the consumer. Solving this problem for each good gives four optimality conditions:

$$r = \alpha_x \frac{Q_x}{K_x} \tag{17}$$

$$w = (1-\alpha_x) \frac{Q_x}{N_x} \tag{18}$$

$$r = \alpha_y \frac{Q_y}{K_y} \tag{19}$$

$$w = (1-\alpha_y) \frac{Q_y}{N_y} \tag{20}$$

The wage rate and rate of return on capital are not indexed by sector. This means that these factors of production are mobile between sectors, and such mobility implies their prices must be the same in the production of each good. The model also has two aggregate conditions which must be met. These state that use of either factor in both sectors cannot exceed the total supply of that factor, or:

$$\bar{K} = K_x + K_y \tag{21}$$

$$\bar{N} = N_x + N_y \tag{22}$$

Where $\bar{K}$ and $\bar{N}$ are the exogenously specified total labor and capital available in the economy. This completes the full specification of model equations. The next step is to bring these equations together using a competitive equilibrium. This begins with a definition for this model.

Definition. *A competitive equilibrium is values for endogenous quantities (K_x, K_y, N_x, N_y, Q_x, Q_y, $\hat{C}$, C_x, and C_y), values for endogenous prices (r, w, $\hat{p}$, p_x, and p_y), given exogenous quantities ($\bar{K}$, $\bar{N}$, Z_x, and Z_y) such that:*

1. *Consumers optimize;*
2. *Firms optimize;*
3. *The labor market clears;*
4. *The capital market clears;*
5. *The consumer's budget constraint is met; and*
6. *One of the goods markets clears.*

To solve for this competitive equilibrium numerically the number of endogenous variables must equal the number of equations. The definition shows there are 14 endogenous variables. The number of equations can be verified by considering each requirement of a competitive equilibrium. If consumers optimize, then the three optimality conditions specified from the consumer's problem must hold. Similarly, if firms optimize each firm has two optimality conditions along with one production functions, for a total of six equations. The labor and capital market clearing conditions add two more equations to the total. Adding the consumer's budget constraint brings the total to 12.

The 13th equation is that one of the goods markets clears, so that $Q_x = C_x$ (alternatively $Q_y = C_y$). Both goods market clearing conditions could be used, but one is redundant when the consumer's budget constraint is included. Instead, a numeraire is specified for the model. Currently, there are not specific units attached to any of the prices/quantities in the model. To do this, set $p_x = 1$ as the 14th equation. This gives 14 equations in 14 unknowns, and the model can be solved if the values of parameters and exogenous variables are specified. The specification of such values is taken up below in the section on social accounting matrices. For comparison, the 14 equations are:

$$C_x = \gamma_c \hat{C} \left(\frac{\hat{p}}{p_x} \right)^{\sigma_c}$$

$$C_y = (1-\gamma_c) \hat{C} \left(\frac{\hat{p}}{p_y} \right)^{\sigma_c}$$

$$\hat{p} = [\gamma_c p_x^{(1-\sigma_c)} + (1-\gamma_c) p_y^{(1-\sigma_c)}]^{\frac{1}{(1-\sigma_c)}}$$

$$r = \alpha_x \frac{Q_x}{K_x}$$

$$w = (1-\alpha_x) \frac{Q_x}{N_x}$$

$$r = \alpha_y \frac{Q_y}{K_y}$$

$$w = (1-\alpha_y) \frac{Q_y}{N_y}$$

$$Q_x = Z_x K_x^{\alpha_x} N_x^{(1-\alpha_x)}$$

$$Q_y = Z_y K_y^{\alpha_y} N_y^{(1-\alpha_y)}$$

$$\bar{K} = K_x + K_y$$

$$\bar{N} = N_x + N_y$$

$$w(N_x + N_y) + r(K_x + K_y) = \hat{p}\hat{C}$$

$$Q_x = C_x$$

$$p_x = 1$$

Open Economy

The majority of the model remains the same if an identically structured region is added. The addition requires specifying any links between the two regions, embodied in both the current and capital accounts. In this section the only links are through the trade of goods, which are produced in both regions. A good produced in one region is similar but not identical to a good produced in another region.

Specifically, the representative consumer (in each region) will now have access to four goods, two from its own producers and two from abroad. This adds another stage to the consumer's optimization problem. As before, the consumer first chooses how much of the composite to consume in the first stage. In the second stage the choice is between a composite of either good, and the composite is made up of domestic and foreign goods. The third stage is a choice between the home or foreign variety of each good.

The first stage is the same as the closed economy case (region one is shown here):

$$\max_{\hat{C}_1} \quad \ln \hat{C}_1$$
$$s.t. \quad w_1(N_{1,x} + N_{1,y}) + r_1(K_{1,x} + K_{1,y}) = M_1 = \hat{p}_1 \hat{C}_1$$

The subscript 1 indicates the first region. The second stage changes from the closed economy case, and now reads (region 1 is shown here):

$$\begin{aligned} \min_{\hat{C}_{1,x}, \hat{C}_{1,y}} \quad & \hat{p}_{1,x}\hat{C}_{1,x} + \hat{p}_{1,y}\hat{C}_{1,y} \\ s.t. \quad & \hat{C}_1 = [\gamma_{1,\hat{c}} \hat{C}_{1,x}^{\rho_{1,\hat{c}}} + (1 - \gamma_{1,\hat{c}}) \hat{C}_{1,y}^{\rho_{1,\hat{c}}}]^{\rho_{1,\hat{c}}} \end{aligned} \tag{23}$$

The notational changes are very important here. The quantities of both goods consumed are now composites of home and foreign goods. Similarly, the parameters have been changed to reflect that fact. Solving this stage still gives three optimality conditions in each region which are necessary for equilibrium.

The third stage in each region consists of two separate optimizations, one for each good. In this stage the consumer in each region decides how much of each good of each region to consume by minimizing total costs (region 1 is shown here for good X):

$$\begin{aligned} \min_{C_{1,x}, C^*_{1,x}} \quad & p_{1,x}C_{1,x} + p^*_{1,x}C^*_{1,x} \\ s.t. \quad & \hat{C_{1,x}} = [\gamma_{1,x,c} C_{1,x}^{\rho_{1,x,c}} + (1 - \gamma_{1,x,c}) C_{1,y}^{\rho_{1,x,c}}]^{\rho_{1,x,c}} \end{aligned} \tag{24}$$

This confusing expression shows that the consumer in region 1 has a choice between good X produced domestically ($C_{1,x}$) or good X produced in the other region but consumed in region 1 ($C^*_{1,x}$). The share parameter ($\gamma_{1,x,c}$) reflects the share of total consumption of good X that comes from domestic production.

This last stage encompasses the famous Armington assumption, and $\sigma_{1,x,c}$ is the so-called Armington elasticity, which quantifies how much the consumer in region 1 is willing to substitute between home and foreign varieties of a good. This is an important parameter in all trade-based models, and is notoriously difficult to pin-down empirically. It is commonly used because it provides a simple way to quantify how much consumers are willing to substitute between products which are similar but produced in different places.

The optimality conditions which come out of this problem are identical in structure to those from stage 2 of the closed economy, but are different notationally. For region one the six optimality conditions from stage 3 are:

$$C_{1,x} = \gamma_{1,x,c} \hat{C}_{1,x} \left(\frac{\hat{p}_{1,x}}{p_{1,x}} \right)^{\sigma_{1,x,c}} \tag{25}$$

$$C^*_{1,x} = (1 - \gamma_{1,x,c})\hat{C}_{1,x} \left(\frac{\hat{p}_{1,x}}{p_x} \right)^{\sigma_{1,x,c}} \tag{26}$$

$$C_{1,y} = \gamma_{1,y,c}\hat{C}_{1,y} \left(\frac{\hat{p}_{1,y}}{p_{1,y}} \right)^{\sigma_{1,y,c}} \tag{27}$$

$$C^*_{1,y} = (1 - \gamma_{1,y,c})\hat{C}_{1,y} \left(\frac{\hat{p}_{1,y}}{p_{1,y}} \right)^{\sigma_{1,y,c}} \tag{28}$$

$$\hat{p}_{1,x} = [\gamma_{1,x,c} p_{1,x}^{(1-\sigma_{1,x,c})} + (1 - \gamma_{1,x,c}) p_{1,x}^{(1-\sigma_{1,x,c})}]^{\frac{1}{(1-\sigma_{1,x,c})}} \tag{29}$$

$$\hat{p}_{1,y} = [\gamma_{1,y,c} p_{1,y}^{(1-\sigma_{1,y,c})} + (1 - \gamma_{1,,y,c}) p_{1,y}^{(1-\sigma_{1,y,c})}]^{\frac{1}{(1-\sigma_{1,y,c})}} \tag{30}$$

The definition of a general equilibrium is similar to the closed economy case, but includes more variables and equations. There are now a total of 40 variables. Before there were 14 equations, which doubles to 28 with both regions, and then sums to 40 when the 6 additional equations from the third stage are added to the system for each region.

Social Accounting Matrix

Numerical solution of any CGE model requires specifying values for the parameters and exogenous variables. The majority of CGE models use calibration to find at least some of these values. Calibration takes a reference data set as an equilibrium, and then calculates the values of a parameter or exogenous variable based on the equilibrium. The standard method is to use input-output data from the national accounts to construct a social accounting matrix (SAM), which serves as the reference data set for an equilibrium. The SAM is then used to back out the values of parameters and exogenous variables.

SAM Basics

The SAM is a record of transactions that take place within an economy during a given year. Specifically, it is an organized matrix representation of all transactions and transfers between different production activities, factors of production, and institutions (households, corporate sector, and government) within the economy and with respect to the rest of the world. A SAM is represented by a square matrix which lists these categories (accounts) on the vertical and horizontal axes. Each category is further disaggregated. For example, the production activities category will distinguish between different goods made by firms, and households is commonly separated between income, consumption, and savings.

The level of aggregation or disaggregation in a SAM is left to the discretion of the modeler, but also

depends on the model being used and data limitations. Each entry into the SAM represents a payment from one account to another. The payee is listed on the horizontal, while the receiver of the payment is on the vertical. This is highlighted in Figure 21, which is a SAM representing the closed economy CGE model outlined above.

	Production Activities		*Factors Production*		*Household*	
	Good X	**Good Y**	**Capital**	**Labor**	**HH**	**Totals**
Good X					100	100
Good Y					100	100
Capital	36	36				72
Labor	64	64				128
HH			72	128		200
Totals	100	100	72	128	200	

Figure 21: Sample SAM for 2-sector, closed economy CGE model

Notice the axis labels on the matrix. Production activities come first, in this model there are two such activities. These are followed by the two factors of production, capital and labor, and then the household. There is no government in the model, nor is there any foreign interaction, so these are not represented in the SAM.

To understand the flows, begin with the first column. This column shows two payments from production of Good X to capital and labor. The payment means that capital and labor are used in production activity X, and the payments represent the compensation for their use.

This compensation to the factors of production is in value terms. One can think of it as price times quantity. So the payment to labor is \$64, and that to capital is \$36. This is why the SAM can be built from the national accounts. One can find production activity X (say toy production) and then read off the compensation to employees in this sector, which is put into the SAM. Complications arise because the aggregation of the SAM is often not the same as that of the underlying national accounts. Also, there are many developing countries where the national accounts are not easily accessible.

Continuing across the columns, production activity Y also pays to capital and labor. Production in this model does not have the use of intermediate inputs, but they can be incorporated by showing the production activity which uses the intermediate input making a payment to the activity from which the good is used. The next column, representing capital, shows that capital income is given directly to the household. The assumption made here, which is standard, is that the household owns the capital and receives any rental payments. The firm or government could own the capital as well if the model was structured in this way. Labor income also goes directly to the household, as shown in the next column. Finally, the last column shows that the household pays \$100 each to consume good X and good Y.

Notice the interdependencies between each column and row. The payments from one account, say production activity X, to another, say capital, is income for the second account. The second account then uses that income to pay some other account, in the case of capital this is payments to the household. And the household then uses that income to pay for goods X and Y. Due to these interactions, there must be consistency between the row and column totals for each account for a SAM to make sense. That is, the income of each account cannot exceed its expenditures. In Figure 21 this is reflected by the fact that each row and column total is the same for each account.

When data is taken from the national accounts this consistency is often lacking. This is because the national accounts have measurement errors or other problems with individual data series. More importantly, the level of aggregation in a CGE model is not at the same level as that of the data in the national accounts, and this will give inconsistent results as well. The SAM must be "balanced" before it can be used as a benchmark for a CGE model. The standard approach is to use a statistical method to achieve this consistency. The most common methods are called RAS and cross-entropy. Unfortunately, both of these will change all of the data in the SAM to achieve consistency and this may have an impact on the results. See Shoven and Whalley (1992) for more on the construction and consistency of social accounting matrices.

Calibration

Constructing a consistent SAM is often the most difficult part of building and simulating a CGE model. Once a balanced SAM is available for such a model, the parameter values can be calibrated. In the closed economy model above, the standard approach would be to use the SAM to derive $\bar{K}$, $\bar{N}$, α_x, α_y, and γ_c. The other parameters and exogenous variables must be specified by the modeler due to the structure of the model. For example, because the share parameter γ_c is based on the SAM, the elasticity of substitution, σ_c, must be input by the modeler. Similarly, allowing α_x and α_y to be derived from the SAM means that Z_x and Z_y must be exogenously specified.

The first step is to assume that all prices are equal to one at the benchmark equilibrium. This is done to ease computation, as the entries into a SAM represent values. With the prices equal to one, the entries then also represent quantities. This gives $\hat{p} = p_x = p_y = w = r = 1$. The $\bar{K}$ and $\bar{N}$ can then be calculated using the SAM. The total amount of capital will be $K_x + K_y = 36 + 36 = 72$. The K_x comes from the payment of production activity X to capital (the value of payments to capital), which is also the amount of capital because $r = 1$. The K_y is also equal to 36 and taken from production activity Y. Similarly, the total amount of labor is calculated based on the SAM and is equal to $64 + 64 = 128$.

Next, the α values can be backed out from the SAM. This is done by rearranging equation (17) to

read:

$$\alpha_x = \frac{rK_x}{Q_x} \tag{31}$$

This equation shows why the Cobb-Douglas parameter α_x is called the capital share. It is the share of production, Q_x, due to payments to capital, rK_x. In this equation the values of r and K_x are known, and the totals category for production activity X gives the value of production (Q_x) in this sector ($= 100$). Once these numbers are plugged in the capital share is equal to 0.36. The same procedure yields a value for α_y.

An alternative option is to specify a value for α_x and to derive a value for the exogenous TFP value, Z_x. This could be done by slightly rewriting equation (31):

$$\alpha_x = \frac{rK_x}{Z_x K_x^{\alpha_x} N_x^{1-\alpha_x}} \tag{32}$$

Given values for α_x this equation can be rearranged to yield a value for Z_x. The remaining parameter is γ_c, which represents the share of total consumption due to the good from production activity X. This comes from using equation (13) and rearranging:

$$\gamma_c = \frac{C_x}{\hat{C}} \left(\frac{\hat{p}}{p_x} \right)^{-\sigma_c} \tag{33}$$

Again, each value in this equation is known or can be read off the SAM. This also shows why either the share parameter (γ_c) or elasticity (σ_c) must be specified by the modeler, as the equation cannot be solved without a value for one or the other. The prices in this equation ($\hat{p}$ and p_x) are both one, C_x is given by the payment of the household to the good from production activity X in the SAM, and $\hat{C}$ is given by the total consumption of the household in the SAM.

While this is a very simple SAM, the basic procedure holds for much larger cases. First pick a year to serve as the base year for the SAM. Then build the SAM to the level of aggregation specified in the model and ensure it is balanced, or consistent. Once this is done, pick elasticity values (or share values) and the values of exogenous values (or specify share values). Finally, use the model equations in combination with the SAM to "calibrate" the remaining parameters or exogenous variables of interest.

2-pd Models

Static models are commonly used for long-run analysis. These models are not able to provide any information on how the dynamics of an economy evolve over time, but rather focus on the starting and ending points. This approach is simpler, as trying to understand the transition between these two points adds substantial conceptual and technical complications. However, there are many interesting

issues where the path is as important as the destination, and this is why dynamic models are popular in macroeconomics. This section outlines two different models which highlight some of the conceptual issues that can arise. In order the show them graphically and simplify the presentation, the models are both two-periods. The first does not include firms, and focuses on consumer decisions alongside the government. The second puts the firms back in the model. Each of these models rely on a general equilibrium framework, and so follow the structure of the previous section by building up from individual components.

A Model of Consumption versus Savings

The consumer problem is still to maximize utility subject to their budget constraint. The difference is that now total utility is summed up over the first and second periods. And because the consumer's actions today impact their options tomorrow, these are necessarily intertwined. To highlight this, we begin with a simple model that only allows the consumer to choose between consumption or savings in the first period, abstracting from the choice between work or leisure. Because there is no labor supply, there is also no firm in this set-up. Some intertemporal tradeoffs can be highlighted in this framework, and important theoretical results such as the permanent income hypothesis and Ricardian Equivalence are illustrated.

The Budget Constraint

Feasible choices for the representative consumer are summarized by the budget constraint. Assume the consumer does not work, but receives an endowment in each period of Y_t. In the first period, this income less lump-sum taxes (T) must be divided between either consumption or savings:

$$C_1 + S_1 = Y_1 - T_1 \tag{34}$$

In this equation C_1 is first-period consumption and S_1 is savings in the first-period. The difference from the static case is that savings is now an option. To allow for such savings, we assume the existence of a financial market where risk-less bonds can be traded, and consumers are able to directly trade these assets without an intermediary. If $S_1 < 0$, the consumer is a lender, and if $S_1 > 0$ the consumer is a borrower.[4]

The risk-less bond is purchased for the price of one consumption good, and pays the lender $(1 + r)$ future consumption goods in the second period. This is the gross rate of return, and r is the real

[4]How can there be borrowing or saving with only one consumer? This part is not explicitly modeled, but there are two ways to think about this. The first is that the consumer trades with those in other countries or regions using the financial market. The other is that this is the budget constraint for only one consumer, and there are many others in the economy to trade with. Either explanation works for our current purposes.

interest rate. Notice that r also gives the tradeoff between current and future consumption. If the consumer is a lender and decides to save, they give up one unit of current consumption goods for $(1+r)$ units of future consumption goods. This makes the price of future consumption in terms of current consumption $\frac{1}{1+r}$, i.e. current consumption per unit of future consumption goods.

In addition to the budget constraint in the first period, the consumer also has a second period budget constraint. Because the model has only two periods, it would not make sense for the consumer to save. For this reason, the second period budget constraint states that income from the endowment and savings is equal to consumption plus taxes, or:

$$C_2 = Y_2 + (1+r)S_1 - T_2 \tag{35}$$

The term on savings, $(1+r)S_1$, can be either positive or negative depending on whether the consumer is a borrower or saver. The fact that the consumer has the option to save creates a link between these budget constraints via S_1. Solve equation (34) for S_1 and then substitute it into equation (35) to derive the intertemporal budget constraint:

$$C_1 + \frac{C_2}{1+r} = Y_1 + \frac{Y_2}{1+r} - T_1 - \frac{T_2}{1+r} \tag{36}$$

This states that the consumer's lifetime consumption must equal their lifetime income less any taxes. The values for the second period are put into present value terms by discounting them using the real interest rate. This is a very powerful equation because it makes clear that the consumer has a restriction over time in addition to each period. To simplify notation, define the consumer's lifetime wealth (we) as:

$$we = Y_1 + \frac{Y_2}{1+r} - T_1 - \frac{T_2}{1+r} \tag{37}$$

Using this simplification, the budget constraint can be rearranged and written in terms of future consumption goods:

$$C_2 = -(1+r)C_1 + we(1+r) \tag{38}$$

This equation gives a tradeoff between current and future consumption. This also looks similar to the static budget constraint, where leisure is replaced by current consumption. Recall that the tradeoff in that case was summarized by the wage rate. Here, the real interest rate gives the tradeoff between current and future consumption. This is very logical, as the benefit from foregoing consumption today is $(1+r)$ units of consumption tomorrow.

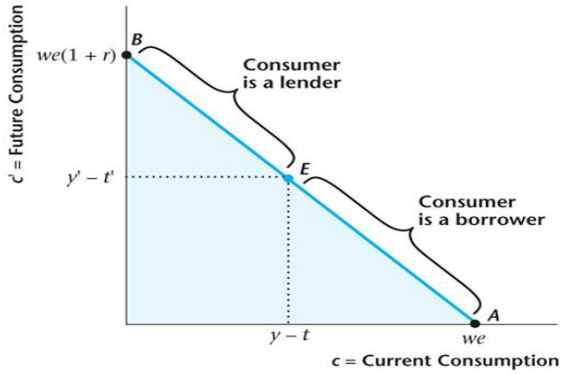


Figure 22: The Consumer's Lifetime Budget Constraint, from Williamson (2011)

The consumer's lifetime budget constraint is plotted in Figure 22. In this case the slope of the constraint is $(1 + r)$, and the intercepts are as shown. Because the consumer can either borrow or lend, there are areas of the budget constraint where the consumer is a borrower and parts where they are a saver. Point E is where the consumer neither borrows nor lends, so that consumption in each period is equal to disposable income (Y_1 - T_1 and Y_2 - T_2). Moving left from point E, consumption in the first period is less than disposable income ($C_1 < Y_1$ - T_1). Projecting this point up to the budget constraint shows that consumption in the second period is then greater than disposable income in that period ($C_2 > Y_2$ - T_2). The opposite is true to the right of point E.

Preferences

The specification of preferences is over consumption in the first or second period instead of work versus leisure. This is a conceptual change, but the mechanics of the indifference curves remain the same. Namely, the three key assumptions made in the static case are also made here. Consumers are assumed to prefer more to less, they like diversity in their consumption bundles, and both current and future consumption are normal goods. The resulting indifference curves are shown in Figure 23

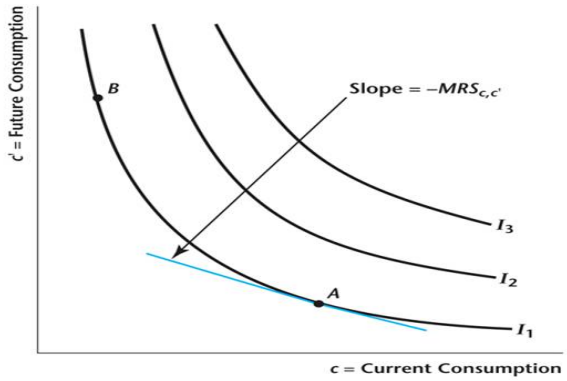


Figure 23: The Consumer's Indifference Curves, from Williamson (2011)

As before, the shape of the curves is a result of the underlying assumptions. Utility increases towards the northeast of the plot, and the convex shape is due to the preference for diversity. In the static case the MRS_{lC} gave the rate at which the consumer was willing to trade leisure for consumption. The slope of these indifference curves is the MRS_{C_2,C_1}, and gives the rate at which the consumer is willing to substitute future consumption for current consumption.

This can be seen by moving down indifference curve I_1. At point B, the slope of the indifference curve is relatively steep, indicating that the consumer is very willing to trade future consumption goods for current ones. To see this, add one unit of current consumption goods by moving to the right of point B. In order to remain on I_1 the consumer would need to give up dC_B future consumption goods. This number is large in absolute value because of the steep slope.

Contrast this with point A, where the MRS_{C_2,C_1} is relatively flat. This means that the consumer is less willing to trade future consumption goods for current ones. As before, begin at point A and move to the right by adding one unit of current consumption goods. In order to stay on I_1 the consumer would need to give up dC_A units of future consumption goods. The fact that dC_B is greater than dC_A means that the consumer is more willing to trade future consumption goods for current ones at point B. This is summarized by saying that the MRS_{C_2,C_1} is higher at point B.

An implication of the convex shape of the indifference curves is that large differences in consumption between the two periods are not desirable. In the example above, at point B the consumer is very willing to give up future consumption for present consumption because future consumption goods are relatively plentiful. At point A they are willing to give up current consumption goods because the reverse is true. The consumption bundles where there is less of a difference in current and future consumption, towards the middle of the indifference curves, are those which are closer to the origin. This indicates they give the same level of utility as points A and B, but are feasible with less income.

Minimizing consumption differences over time in this way is called consumption smoothing. When put in this light, assuming a preference for diversity in current and future consumption is a sensible assumption. Most people seem to prefer relatively stable levels of consumption over time instead of large differences from period to period.

Consumer Optimization

The optimal consumption bundle is the point where the indifference curve just touches the budget constraint. The consumer wants to be on the budget constraint because this will touch the highest indifference curve possible. Figure 24 shows the optimal bundle is at point A.

If the consumer did not choose point A, they could increase utility through trade. Take for example point E, and imagine there is an indifference curve which intersects the budget constraint at this point. The slope of this indifference curve would be less than the slope of the budget constraint. That is, the consumer values future consumption more than the market. Instead of choosing this point, the

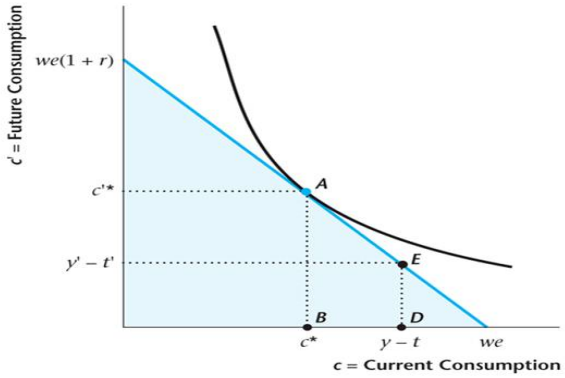


Figure 24: The Consumer's Optimal Bundle, from Williamson (2011)

consumer could give up a unit of current consumption for a unit of future consumption. The consumer would be willing to accept MRS_{C_2,C_1}, but the market would be willing to pay $(1+r)$. Because we know that $MRS_{C_2,C_1} < (1+r)$ (the slope of the indifference curve is less than the slope of the budget constraint), the consumer is willing to make this trade. And the trade can be made in the market because those on the other side receive current consumption, which they value more than future consumption. Only at point A can no-one be made better off, and this is the optimal bundle. The optimality condition in this case is $MRS_{C_2,C_1} = (1+r)$.

Changes in Income

Changes in income will vary the optimal bundle for the consumer, but the impact depends on the specific details of the respective change, and whether the consumer is a lender or a borrower. For purposes of brevity, all of the experiments below view the consumer as a lender.

Consider first an increase in current period income. This is a temporary increase in income for the consumer in the first period, where income in the second period does not change. This is similar to the income effect in the static case as shown in Figure 25.

The initial optimal bundle is at point A. An increase in income will shift out the budget constraint as shown because lifetime wealth increases. This can be seen from the definition of wealth:

$$we = Y_1 + \frac{Y_2}{1+r} - T_1 - \frac{T_2}{1+r}$$

In this case Y_1 increases to Y_1^*, so that we increases to we^*, which is represented as we_2 in the figure. The higher level of income makes additional consumption bundles feasible for the consumer, which allows them to move to a higher indifference curve. The new optimal bundle will be something like point B. Notice that the increase in current consumption is less than the increase in income. Current income increases by the difference between E_1 and E_2, but current consumption increases by the

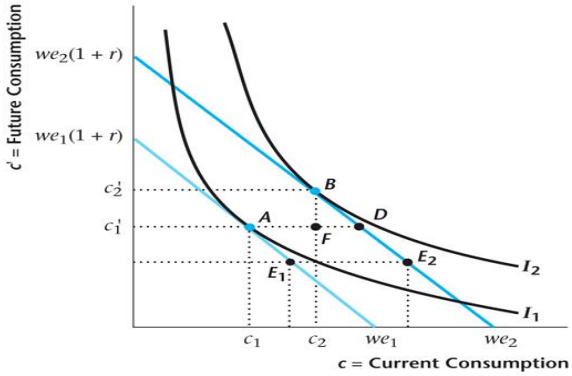


Figure 25: The Effects of an Increase in Current Period Income, from Williamson (2011)

amount between A and F, which is smaller.

The reason for this is consumption smoothing, or the fact that the consumer prefers diversity in their consumption bundle. Faced with a temporary increase in income, the consumer chooses to spread the benefits out over time instead of taking all of the gains in the first period. This can be seen in the increase of future consumption from F to B. The total impact of a temporary increase in income will be to increase income in both periods, and Figure 25 shows that point B is to the right of and above point A. One other implication is that the consumer's savings must increase with this temporary increase in income. This is because the rise in income is greater than the rise in current consumption.

An increase in future income will have the same effect as shown in Figure 26. In this case future wealth rises, which shifts out the budget constraint. It is optimal for the consumer to smooth income as before. In this case savings must fall for the consumer to increase current consumption, the opposite of what happened when the increase in income came in the current period. The caveat to this analysis is that the consumer can perfectly predict an increase in future income. This comes from the perfect foresight assumption used in this two period model.

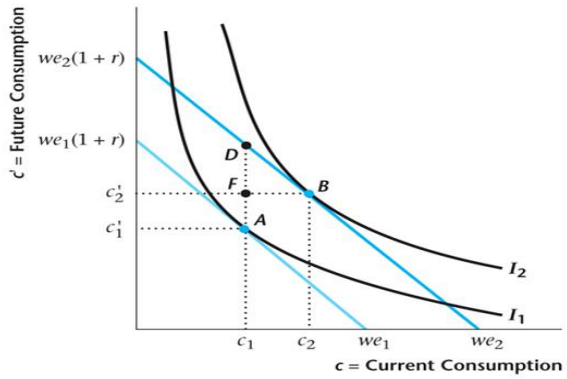


Figure 26: The Effects of an Increase in Future Income, from Williamson (2011)

Each of these first two experiments are temporary increases in income. Consider next an increase in both future and current income, or a permanent increase in income. The impact of such a change in current consumption is much larger, as shown in Figure 27.

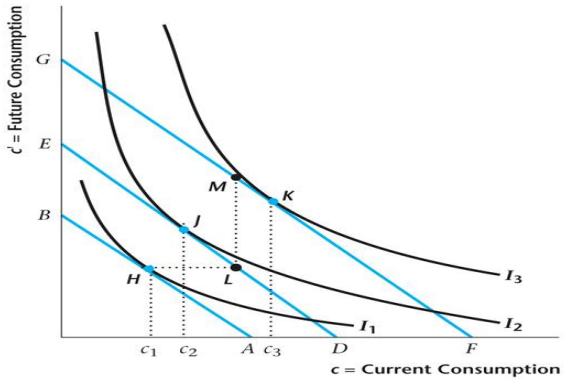


Figure 27: The Effects of an Increase in Permanent Income, from Williamson (2011)

In this case the budget constraint shifts from BA to ED due to the rise in current income, and from ED to GF because of the increase in future income. Both current and future consumption rise more than with a temporary increase in income. The policy implications of this experiment follow from the fact that current consumption rises by more with a permanent increase in income than a temporary one.

This difference in permanent versus temporary impacts is an example of the permanent income hypothesis. According to this hypothesis, the primary determinant of a consumer's current consumption is his or her permanent income, not their current income. In this model permanent income is the lifetime wealth of the consumer. The logic follows straight from the fact that consumers do not like large changes in consumption over time. To avoid these changes any temporary gains in income will smoothed over time, with current consumption rising by less than current income.

Figure 27 shows this is not necessarily true with permanent changes in income. The final optimal bundle at point K may lead to a rise in current consumption which is larger than the rise in current income (but not total income). Ultimately this depends on the preferences of the consumer, but the possibility exists when there is a permanent increase in income, unlike the temporary case.

Changes in the Real Interest Rate

Another experiment which can be conducted is to vary the real interest rate on savings, r. Because we are assuming the consumer is a lender, this will impact their income. And because it is the opportunity cost of consumption, it will also impact the savings/consumption decision of consumers. As with a change in the wage rate in the static case, this will induce both income and substitution effects. These are shown in Figure 28 for an increase in the real interest rate.

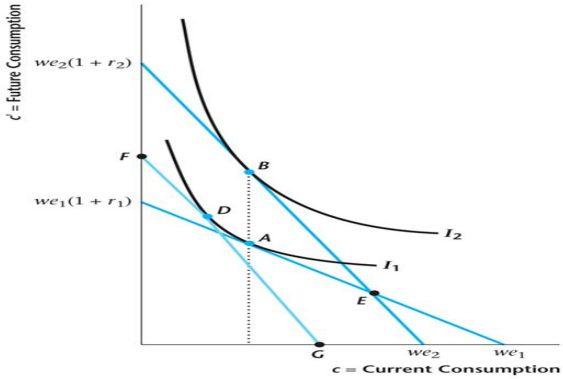


Figure 28: The Effects of an Increase in the Real Interest Rate, from Williamson (2011)

A rise in the real interest rate will increase the slope of the budget constraint. For the lender, this raises the opportunity cost of consumption today (or the price of current consumption). The consumer will therefore choose to substitute away from current consumption towards future consumption. In Figure 28 this is shown by rotating the original budget to line FG. At point D the slope of FG is equal to the new real interest rate and is just tangent to the original indifference curve. This is how much the consumer would substitute away from current consumption towards future consumption to keep the same level of utility. This is called the intertemporal substitution effect.

This is not the end of the story. The higher real interest rate also makes the consumer wealthier, all else equal. This shifts out the budget constraint and induces an income effect. Because current and future consumption are normal goods, both increase. The final optimal bundle is at point B in Figure 28. In total, future consumption must increase because both income and substitution effects work in the same direction. There are offsetting forces on current consumption, however, and the total impact is ambiguous.

Government

The government in this model generates income through lump sum taxes and by issuing bonds. It then spends these by purchasing consumption goods. As with the consumer, the government has a budget constraint for each period. In the first period, it can only spend what is generated in taxes or borrowed (or lent if the government has a surplus):

$$G_1 = T_1 + B_1 \tag{39}$$

In the second period the government can only spend what it generates in tax revenue less the value of

the earlier borrowing:

$$G_2 = T_2 - (1+r)B_2 \tag{40}$$

As with the consumer these can be combined to generate the government's intertemporal budget constraint:

$$G_1 + \frac{G_2}{1+r} = T_1 + \frac{T_2}{1+r} \tag{41}$$

This is an important equation. Notice that the bonds have disappeared from the intertemporal budget constraint. This is because they must be paid back, so that the government can only spend what it receives in tax revenue over both periods. This constraint makes clear that a government which does not default must generate its revenue in taxes.[5]

Competitive Equilibrium

As with the static models, a competitive equilibrium is used to bring the model together. The definition changes slightly when the time dimension is added, but the basic ingredients are the same.

Definition. *A competitive equilibrium in the current model is the values of endogenous quantities (C_1, C_2, T_1, T_2, B_1, and B_2) and an endogenous price (r), given exogenous quantities (Y_1, Y_2, G_1, and G_2) such that the following conditions are satisfied:*

1. *The consumer optimizes each period given market prices;*
2. *The credit market clears; and*
3. *The government budget constraint is satisfied each period.*

The largest difference between this intertemporal equilibrium and the static one is that the budget and government constraints must hold each period. Because the consumer can save, the market for credit must also clear as well. The condition for this is that the value of bonds issued must be equal to the savings of the consumer in the first period. As a final point, this definition also implies a fourth condition holds, namely that the goods market clears in either period $Y_t = C_t + G_t$.

Ricardian Equivalence

The model in its current form can be used to illustrate the concept of Ricardian equivalence. This theorem views the timing of taxes by government as neutral, in that varying taxes does not change the

[5] This statement is true in the context of the current model. However, if there was money in the model the government could reduce its real debt burden by manufacturing inflation. Generating revenue in this manner is termed seigniorage.

real interest rate or savings/spending decisions of consumers. Ricardian equivalence is a very powerful result that states that the method of financing current expenditures (current taxes or future taxes) does not matter. There are many qualifications to this result, primarily having to do with the limitations of the model, but it is a valuable starting point to think about the impact of tax cuts.

To show that Ricardian equivalence holds in this model, begin with the government's intertemporal budget constraint, equation (41):

$$G_1 + \frac{G_2}{1+r} = T_1 + \frac{T_2}{1+r}$$

This equation shows that any changes in expenditures must be matched by corresponding changes in taxes in either period. While it is true that current expenditures may be financed by borrowing and issuing bonds, these bonds can only be paid back through taxation. The government's intertemporal budget constraint makes this point clear. This equation can then be substituted into the consumer's intertemporal budget constraint, equation (36):

$$C_1 + \frac{C_2}{1+r} = Y_1 + \frac{Y_2}{1+r} - T_1 - \frac{T_2}{1+r}$$

which gives:

$$C_1 + \frac{C_2}{1+r} = Y_1 + \frac{Y_2}{1+r} - G_1 - \frac{G_2}{1+r} \tag{42}$$

Now consider a change in the method of financing a given level of government spending. Specifically, assume the economy is in an equilibrium with specified levels of r, C_1, C_2, T_1, T_2, G_1, and G_2. Suppose the government decides to decrease T_1 by ΔT without changing G_1 or G_2. According to the government's intertemporal budget constraint, future taxes must rise by $\frac{\Delta T}{1+r}$. Notice that nothing else has been changed save the timing of taxes, whereby there is a current tax cut and a future tax rise.

Equation (42) shows that nothing changes for the consumer. The right-hand side is still the same because government spending has not changed, and neither has income. The real interest rate does not change either. This is because the increased government borrowing due to the tax cut is matched by the increased savings of the consumer, so that the price which clears the credit market (the real interest rate) is unchanged. This occurs because consumers save the tax cut, as they know that this will need to be repaid in the future. The result is that the timing of taxes does not matter, or that the Ricardian equivalence theorem holds.

The Ricardian equivalence theorem is an important theoretical result, but is unlikely to hold perfectly. The first issue is that consumers are assumed to care as much about the future as about the present. It might be the case that taxes are deferred so long that future generations will be stuck with the bill,

so that current consumers do not believe they will have to repay the borrowing. It is also assumed that taxes change by the same amount for all consumers. If some consumers do not have to repay the tax cut, they may well choose to change their consumption pattern, as this is an income effect for them. A final important assumption is that there are perfect credit markets in that consumers can borrow and lend as much as they please. If consumers are credit-constrained, meaning they cannot borrow as much as they like, they may view a tax cut as a loan and change their consumption profile to better match their preferences.

These are significant caveats to the Ricardian equivalence theorem. Still, the thrust of the theorem remains valid: current changes in taxes have consequences for future taxes.

The Full Model

The simplified two-period model of consumption versus savings was introduced to focus on the consumer's intertemporal choice between savings and spending. This section brings back the choice between work and leisure for the consumer. Instead of getting an exogenous endowment each period, the consumer generates income by supplying labor to the firm, and also receives profits from the firm each period as well. To summarize consumer behavior a labor supply curve and a demand curve for current consumption goods are constructed. The firm's problem changes from the static case as well. The firm chooses how much labor to demand each period, but also how much to invest in growing the capital stock. These two decisions of the firm can be summarized in a labor demand curve and an optimal investment schedule. The government's intertemporal budget constraint is the same as before, and the model is closed by specifying a competitive equilibrium.

Consumers

The consumer's budget constraint brings together those from the static and simplified two-period cases. In the first period, the consumer can choose to consume or save their income. This income comes from supplying labor to the firm or from any of the firm's profits, less lump-sum taxes:

$$C_1 + B_1 = w_1(h_1 - l) + \pi_1 - T_1 \tag{43}$$

The variable definitions are the same as in previous sections. The second period budget constraint is similar, except that the consumer will not save:

$$C_2 = w_2(h_2 - l) + (1 + r)B_1 + \pi_2 - T_2 \tag{44}$$

Second period consumption is financed by income less taxes plus any benefits or costs of savings/borrowing in the initial period. Combining these two equations by substituting out B_1 gives

the consumer's intertemporal budget constraint:

$$C_1 + \frac{C_2}{1+r} = w_1(h_1 - l) + \frac{w_2(h_2 - l)}{1+r} + \pi_1 - T_1 + \frac{\pi_2 - T_2}{1+r} \tag{45}$$

In present value terms, consumption over both periods is limited by lifetime wealth.

In the full model consumer preferences cannot be specified graphically, but may be characterized in four optimality conditions. These conditions must be met for the choices of the consumer to be optimal. The fact that there are four such conditions is the reason graphical representation is not possible. In contrast, the static case was one dimensional in that there was only a choice between consumption or leisure. The same was true of the simple two period case, as the choice was between consumption in the first or second period. However, the same assumptions hold with respect to these multi-dimensional preferences, namely that more is preferred to less, diversity in each is preferred, and each of the goods is normal.

The four optimality conditions are with respect to the choice between consumption versus leisure in the first period, consumption versus leisure in the second period, consumption in the first period versus consumption in the second period, and leisure in the first period versus leisure in the second period. The decision between consumption and leisure in the first period stays the same as the static case. The optimal point is where the consumer supplies labor until the marginal rate of substitution between leisure and consumption equals the wage rate:

$$MRS_{l_1 C_1} = w_1 \tag{46}$$

The same condition holds in the second period as well between leisure and consumption:

$$MRS_{l_2 C_2} = w_2 \tag{47}$$

The intertemporal consumption choice remains the same also, as the consumer equates the marginal rate of substitution in consumption between the first and second periods to the real interest rate:

$$MRS_{C_2 C_1} = 1 + r \tag{48}$$

The final condition combines each of these and specifies the optimal intertemporal substitution of leisure. This conditions states that the consumer should supply labor until the point where the marginal rate of substitution between leisure in the first and second period (MRS_{l_2,l_1}) equals the ratio

of wage rates in present value terms:

$$MRS_{l_2 l_1} = \frac{w_1(1+r)}{w_2} \tag{49}$$

The right-hand side expresses the relative price of leisure over time. If the current wage rate or the real interest rate rises, optimality requires the MRS_{l_2,l_1} to rise as well. This means that the value of future leisure must fall relative to current leisure, so that the consumer substitutes leisure from the first period to the second. This leads to increased labor supply in the first period. The opposite holds with a rise in the second period wage or a fall in the first period wage or the real interest rate.

Although it is not possible to show the preferences graphically, the implied behavior from the optimality conditions can be summarized in two plots. The first of these is the current labor supply curve, which relates the wage rate in the first period to labor supply. From the static model we know that the labor supply of the consumer rises with the wage rate in the current period. This assumes that the substitution effect dominates the income effect of a wage rate increase, and is shown in Figure 29

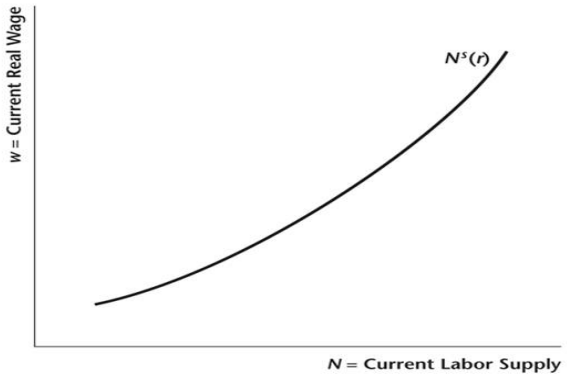


Figure 29: The Labor Supply Curve in the Two-period Model, from Williamson (2011)

This curve shifts when the consumer's lifetime wealth changes. An increase in the present value of wealth leads to a positive income effect, whereby the consumer demands more current leisure and supplies more labor in the first period, shifting the labor supply curve to the left. Figure 30 shows such an increase in lifetime wealth on the labor supply curve. A rise in the amount of lump-sum taxes (in either period) will have the opposite impact.

The fact that the consumer has an optimality condition which characterizes the intertemporal substitution of leisure means that the labor supply curve also depends on the real interest rate. As was explained above, an increase in the real interest rate leads to increased labor supply in the current period, which shifts the labor supply curve right as shown in Figure 31.

The second summary graph is the consumer's demand for consumption goods in the first period versus aggregate income. As aggregate income rises the consumer's demand for current consumption does so

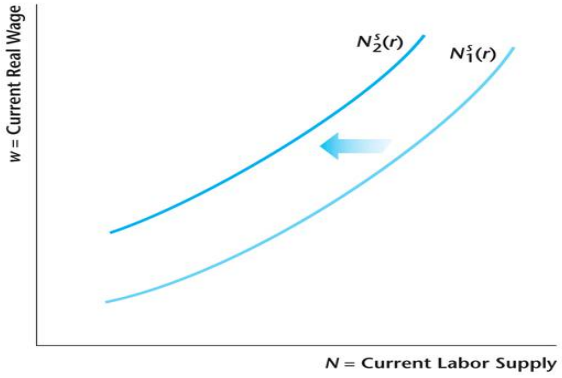


Figure 30: The Impact of an Increase in Wealth on Labor Supply, from Williamson (2011)

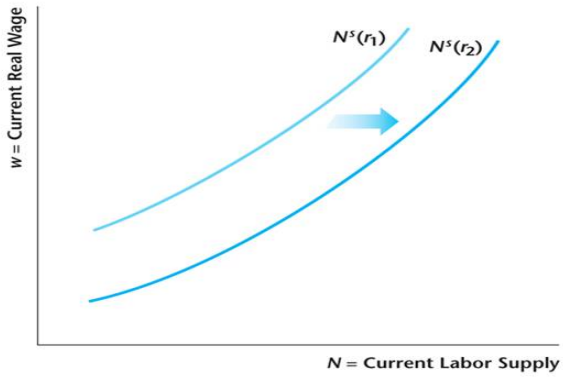


Figure 31: The Impact of an Increase in the Real Interest Rate on Labor Supply, from Williamson (2011)

as well because these goods are normal. However, as the consumer accumulates more income the addition of another unit of income leads to increases of less and less in terms of demand for first period consumption goods. This reflects the consumer's preferences for diversity. The curve is shown in Figure 32.

The slope of this demand curve is called the marginal propensity to consume (MPC). This is the amount of increase in current consumption with a one unit increase in aggregate income. We should expect that as there is more and more income that the MPC falls as shown in Figure 32 due to a preference for diversity. As with the labor supply curve, changes in either lifetime wealth or the real interest rate will shift this curve. In the case of an increase in lifetime wealth, maybe due to a reduction in taxes or an increase in future income, the demand for current consumption goods will increase at any level of aggregate current income, shifting the curve upwards as shown in Figure 33.

A rise in the real interest will have the opposite impact, as shown in Figure 34. This is because the real interest rate is also the price of current consumption, and when it rises the consumer shifts consumption from the current period to the future due to the substitution effect. If the consumer is a

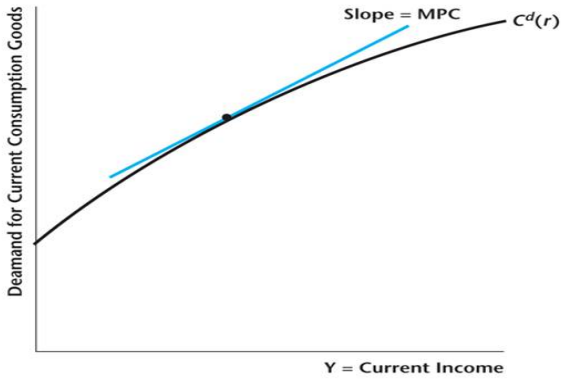


Figure 32: The Demand for Current Consumption Goods, from Williamson (2011)

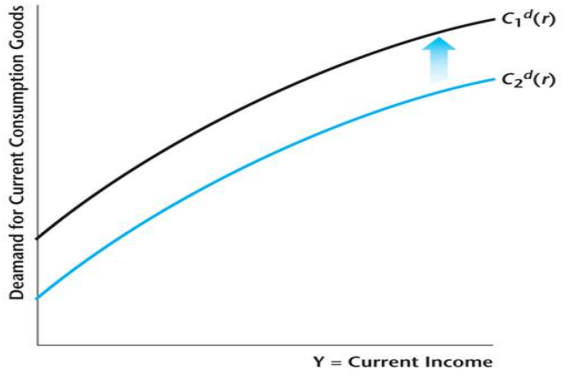


Figure 33: The Impact of a Rise in Lifetime Wealth on the Demand for Current Consumption Goods, from Williamson (2011)

saver, there is also an income effect which raises current consumption due to greater lifetime wealth. We assume the substitution effect dominates, so that the curve shifts downward when r rises.

This demand for current consumption goods is only a part of the total demand for current goods. The firm also demands such goods for building the capital stock (investment), and the government purchases these goods as well.

Firms

The firm produces consumption goods in either period using a production technology which depends on capital and labor. In the first period, the firm must decide how much labor to hire for use in the production process. Although the first period capital stock is fixed, the firm must also decide how much to invest in the construction of new capital for use in the second period. This is called investment, and it is assumed that the firm can costlessly convert one unit of the current consumption good to one unit of an investment good. We assume that the firm finances its investment out of profits and does not issue bonds.

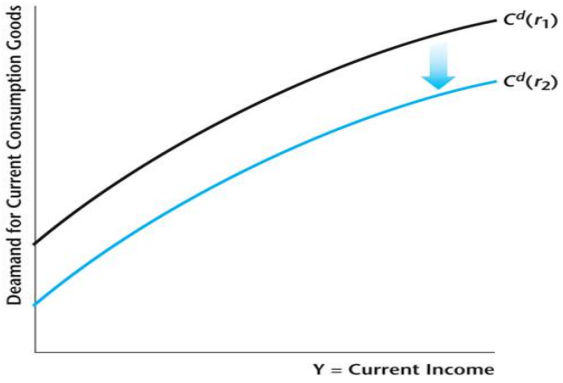


Figure 34: The Impact of a Rise in the Real Interest Rate on the Demand for Current Consumption Goods, from Williamson (2011)

In the second period the firm again chooses how much labor to demand for production. There is no investment in this period, however, as it is not optimal to leave capital stock which will be unused. The firm is able to costless convert this capital stock to second period consumption goods after production has taken place, and these goods can then be sold.

The firms problem is then to choose their optimal labor demand in both periods and first-period investment to maximize profits. Profit in the first period (π_1) is total production less labor costs less any costs for investment, or:

$$\pi_1 = Y_1 - w_1 N_1 - I_1 \tag{50}$$

The capital stock grows due this investment, but also depreciates over time. This is shown by the capital accumulation equation:

$$K_2 = (1-d)K_1 + I_1 \tag{51}$$

This equation shows that the capital stock in the second period (K_2) equals investment in the first period (I_1) plus the undepreciated capital remaining from the first period. The d represents the depreciation rate of capital, or the rate at which capital wears out. In the second period $I_2 = 0$ and it is also optimal for $K_2 = 0$ after production has taken place, so that profit is production less labor costs plus the sale of any undepreciated capital:

$$\pi_2 = Y_2 - w_2 N_2 + (1-d)K_2 \tag{52}$$

As specified, the firm's labor demand problem does not have an intertemporal dimension in either

period. This is because labor demand in each period impacts only the current period's profits, so that the optimality condition derived for the static case holds in either period: $MRS_{l_t c_t} = w_t$. Recall that this means the labor demand curve coincides with the marginal product of labor, as shown in Figure 35.

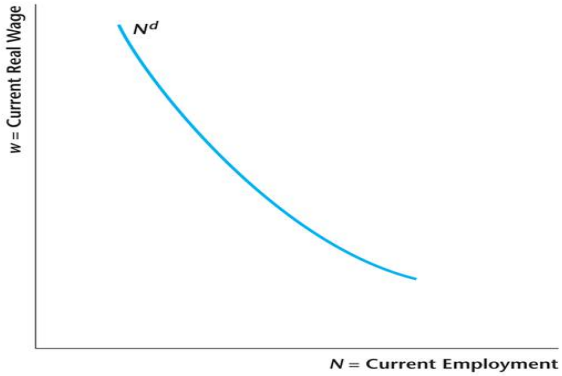


Figure 35: The Firm's Labor Demand Curve, from Williamson (2011)

As in the static case, an increase in current TFP increases the marginal productivity of labor, which shifts the curve to the right as shown in Figure 36. Increasing the current capital stock also raises the marginal productivity of labor, implying a higher wage rate from the optimality condition, so that this also shifts the curve right as shown in Figure 36.

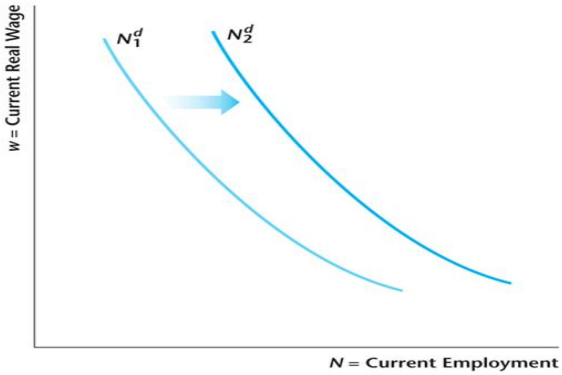


Figure 36: Impacts of Increases in Current TFP or the Current Capital Stock on the Labor Demand Curve, from Williamson (2011)

In contrast, the firm's investment decision is intertemporal, and weighs the costs of such investment against the benefits. The optimal point is where the marginal costs of investment equal its marginal benefits. The marginal cost of investment is the loss of current profits due to an extra unit of investment. Equation (50) shows that if I_1 rises by one unit then π_1 falls by one unit, meaning that the cost is 1.

The benefit of this extra unit of investment shows up in the second period as future profits and is given

by equation (52). The additional investment increases the capital stock by one unit, so that K_2 is one unit greater, implying a gain in profits of $(1-d)$. The extra unit of capital also increases production in the second period, as capital is a direct input into the production function. The extra output due to one unit rise in the capital stock is given by the marginal product of capital in the second period, or MP_{K_2}. Equating the marginal benefits to the marginal costs (both in present value terms) gives:

$$\frac{MP_{K_2}+1-d}{1+r}=1 \tag{53}$$

The left-hand side is the marginal benefit of a unit of investment in present value terms, and the right-hand side is the marginal cost. This can be rearranged to give:

$$MP_{K_2}-d=r \tag{54}$$

Equation (54) is the firm's optimality condition for investment. It states that the firm should investment until the benefit from that investment is equal to the cost, as explained above. One could also use an arbitrage argument to motivate this condition. This is because the firm has the option of purchasing government bonds instead of investing in capital stock. It only makes sense to invest in capital stock if the return is greater than or equal to that which could be generated by purchasing government bonds. Thus the rate of return on investment in capital $(MP_{K_2}-d)$ must just equal or be slightly greater than the rate of return on bonds (r). This relationship is plotted in Figure 37.

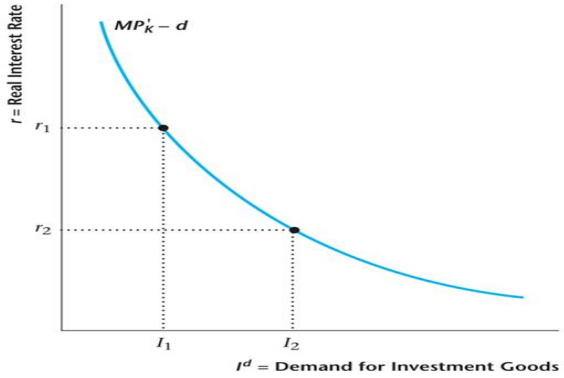


Figure 37: The Firm's Optimal Investment Schedule, from Williamson (2011)

There is an inverse relation between I and r implied by the optimal investment rule. As the real interest rate falls, the marginal product of capital must fall to maintain optimality (d is constant). Intuitively, the lower r makes investment in capital more attractive relative to bonds, and this increased investment in the capital stock works to lower the future marginal product, which gives the inverse relationship.

The curve can shift due to changes in the marginal product of capital. In this model, the future level of TFP and the current capital stock can both change the future marginal product of capital. The curve shifts out with a rise in future TFP because the marginal product of capital in the second period is higher at any given r. Because r has not changed investment in capital is more attractive than purchasing bonds, leading to the shift shown Figure 38. This same shift occurs if there is a fall in the current capital stock, possibly due to a war or natural disaster. A lower capital stock corresponds to a higher marginal product, again making investment in capital preferable to investment in government bonds.

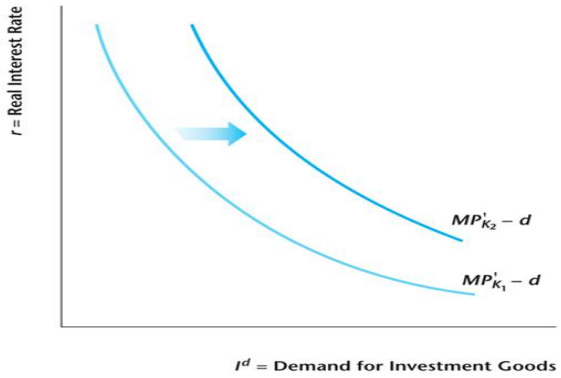


Figure 38: The Impact of an Increase in Future TFP or a Decrease in Current Capital to the Firm's Optimal Investment Schedule, from Williamson (2011)

Government

The government can tax the consumer each period and issue or purchase bonds. Combining the two constraints as above yields the government's intertemporal budget constraint:

$$G_1 + \frac{G_2}{1+r} = T_1 + \frac{T_2}{1+r} \tag{55}$$

Competitive Equilibrium

As with the other models, a competitive equilibrium is used to bring the model together.

Definition. *A competitive equilibrium in the full two-period model is the values of endogenous quantities (Y_1, Y_2, C_1, C_2, N_1, N_2, K_2, I_1, T_1, T_2, B_1, and B_2) and endogenous prices (r, w_1, and w_2), given exogenous quantities (G_1, G_2, K_1, Z_1, and Z_2) such that the following conditions are satisfied:*

1. *The consumer optimizes each period given market prices;*
2. *The firm optimizes each period given market prices;*

3. *The labor market clears each period;*

4. *The credit market clears;*

5. *The government budget constraint is satisfied each period; and*

6. *The current goods market clears.*

These five conditions imply that the goods market clears in the second period.

Graphical Representation

In its current form the model is a system of equations which characterize the optimal decisions for consumers and firms, and ensure all markets clear as specified in the definition of a competitive equilibrium. The value of using a two period model is that the competitive equilibrium can be shown graphically, and this makes experimenting with the model relatively easy. The complete model consists of two different graphs, one summarizes the current labor market and the other the supply and demand for current goods. The discussion here focuses on the current period, but similar plots can be constructed for the second period as well.

The current labor market is summarized through the use of labor supply and labor demand curves, both of which were derived above. The labor demand curve comes from the optimality condition of the firm, and traces the points where the wage rate equals the marginal product of labor. The labor supply curve is derived from the consumer's optimality conditions. For a given real interest rate, this curve traces the points where a given supply of labor results in a value for the marginal rate of substitution between current leisure and consumption which equals the wage rate (i.e. $MRS_{l,C} = w$). A change in the real interest rate will shift this curve because it changes the consumer's decision on how to substitute leisure over time. The intersection of these two curves yields an equilibrium wage and labor supply for the current period.

The second curve summarizes the relationship between the real interest rate and aggregate output. It is comprised of an output supply curve and an output demand curve, which respectively summarize the production and demand for final goods. The output supply curve is derived beginning with equilibrium in the labor market, which gives the labor input to production. Because current capital is fixed, production depends on this labor input and the level of TFP given the capital stock. Figure 39 shows graphically how the level of production can be backed out from the equilibrium labor supply given the level of TFP and capital stock.

Because the labor supply curve is drawn for a particular real interest rate, there is also an implied relationship between the real interest rate and final goods output. To see this, consider an increase in the real interest rate as shown in Figure 40. This will result in increased labor supply at any given wage rate, raising the equilibrium level of employment. Because employment is higher, the firm can

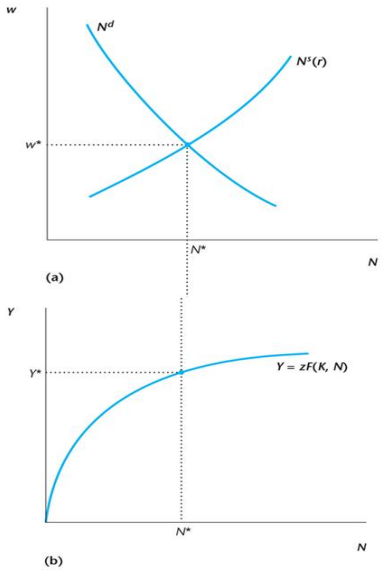


Figure 39: Deduction of Final Goods Production Using Labor Market Equilibrium, from Williamson (2011)

produce additional goods, resulting in a higher level of output. The upward sloping output supply curve shown in Figure 40 summarizes this relationship between r and Y.

This curve should be interpreted as the output/real interest pairs for which the labor market is in equilibrium. Because the output supply curve depends on the labor supply curve and the production function, it shifts with either of these two curves. Take for example an increase in current government spending as shown in Figure 41. This elevated spending is reflected through the consumer's budget constraint as an increase in taxes, or a decrease in lifetime wealth as was explained above. The reduction in lifetime wealth induces the consumer to decrease their leisure because it is a normal good, and this moves out the labor supply curve, as shown in the figure. The new equilibrium labor supply, at the original real interest rate, gives a higher level of output. Because the real interest rate is the same, but output is higher due to the labor supply response from the additional government spending, the output supply curve shifts to the right as in Figure 41.

A similar rightward movement in the output supply curve occurs current TFP rises, as shown in Figure 42. In this case, the elevated level of TFP raises the marginal product of labor at any amount of labor supply. The demand for labor then shifts to the right because each worker can produce more. The result is a higher level of employment in equilibrium with the original real interest rate, and this moves out the output supply curve as shown in the figure.

The full model is completed with construction of the output demand curve. This curve plots combinations of the real interest rate and total demand for final goods which are consistent with competitive market equilibrium. Total demand for current goods $[Y_1^d]$ is the sum of the demand from

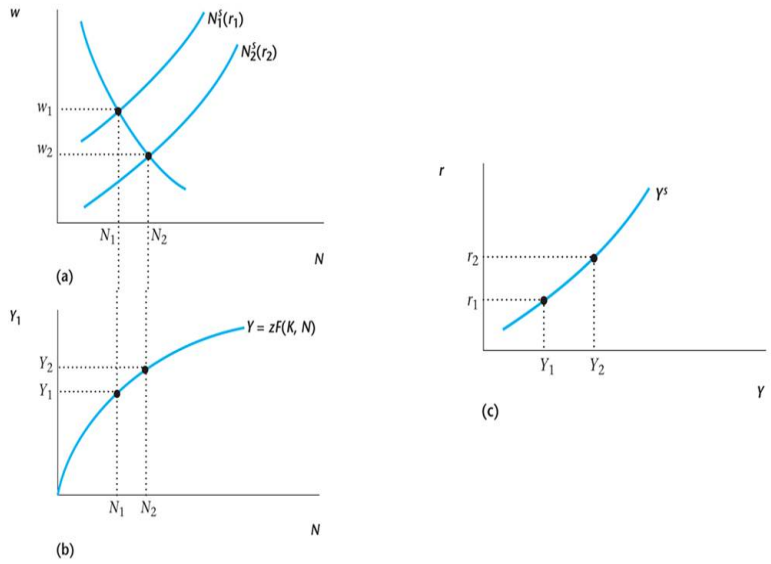


Figure 40: Construction of the Output Supply Curve, from Williamson (2011)

consumers $[C_1^d(Y_1^d, r)]$, the demand for investment goods by firms $[I_1^d(r)]$, and government purchases of goods $[G_1]$, or $Y_1^d = C_1^d(Y_1^d, r) + I_1^d(r) + G_1$. Because both consumption and investment depend on the real interest rate, so does the total demand for current goods. This means that the output demand curve can be constructed in the same way as the output supply curve, by varying the real interest rate and seeing how total current demand changes.

Total current demand can be used instead of production or income, because in equilibrium it must be the case that total current demand is the same as total current supply which is the same as total current income. This can be shown graphically by plotting Y_1^d against Y_1 as is done in Figure 43. The 45 degree line summarizes all points where $Y_1^d = Y_1$, so any value for these variables consistent with equilibrium must lie on this line.

The line which summarizes total demand for current goods has a concave shape as income rises. This is because the only variable in the expression which changes with income is current consumption. Increases in current consumption with additional income get lower and lower as income rises. This is another way of saying that the concave shape reflects the fact that the consumer has a declining marginal propensity to consume additional income. The intersection of these two lines gives the equilibrium level of current demand/income at a particular real interest rate.

The output demand curve is constructed by varying this real interest rate to see how current demand/income change as shown in Figure 44. Consider an increase to the real interest rate. This will make savings relatively more attractive for the consumer, reducing current consumption. Similarly, a higher real interest rate makes investment less attractive for firms, lowering current investment. These

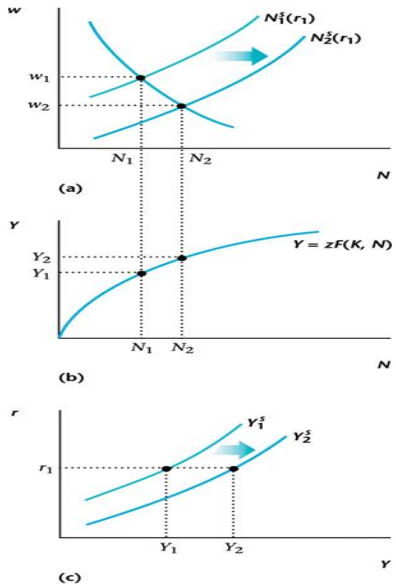


Figure 41: The Effect of an Increase in Government Spending on the Output Supply Curve, from Williamson (2011)

two effects work to reduce current demand, which shifts down the curve summarizing total current demand. The result is a lower level of current demand/output, which is reflected in the output demand curve shown in Figure 44.

As with the output supply curve, there are various factors which can shift the output demand curve. The same change in government spending shown above results in an increase to the demand for current goods, shifting this curve up as shown in Figure 45. The equilibrium level of output is higher and the real interest rate has not changed, resulting in a movement out of the demand curve as shown in the figure. Other factors which shift the output demand curve include changes in the present value of taxes, a variation in future income, movements in future TFP, and changes in the current capital stock.

Figure 46 brings the model together. These two graphs, one of the labor market and the other of the current goods market, show this model's competitive equilibrium.

Experiments

At this point the model can be used for various experiments. Here, changes to future TFP and credit market uncertainty are considered. Other possible examples include changes in current of future government spending, higher or lower current TFP, and increases or destruction of current or future capital stock.

The first experiment considered is an increase in future total factor productivity, possibly due to a technological breakthrough. The initial impact of higher future TFP is to raise the future marginal

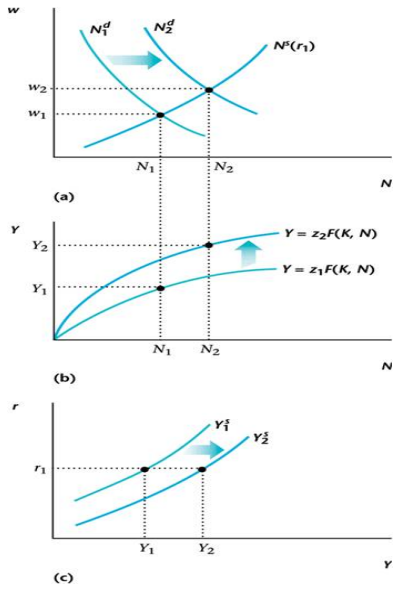


Figure 42: The Effect of an Increase in TFP on the Output Supply Curve, from Williamson (2011)

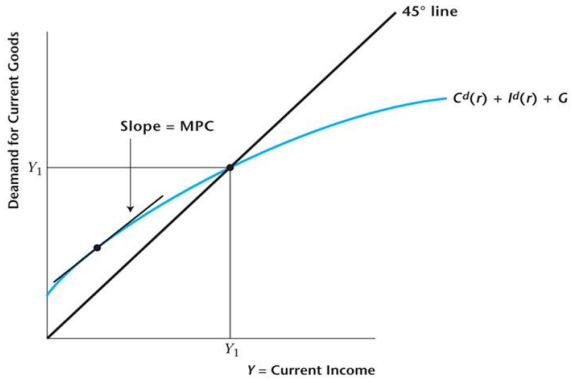


Figure 43: The Demand for Current Goods, from Williamson (2011)

product of capital, as any capital in the next period will produce more. This leads the firm to increase current investment, as it shifts out the optimal investment schedule. Panel (b) of Figure 47 shows that this results in a right-ward movement in the output demand curve.

In equilibrium both the real interest rate and level of output are higher, each of which has feedback effects. The higher real interest rate leads the consumer to supply additional labor, moving the labor supply curve to the right as shown in Panel (a) of Figure 47. But the higher income has an income effect that works in the other direction on the labor supply curve. The new equilibrium wage and employment level depend on which effect dominates. Assuming that the change in the real interest rate dominates, the equilibrium level of employment is higher and the wage rate lower. Each of these also have feedback effects back to the current goods market, but these are not shown in the figure.

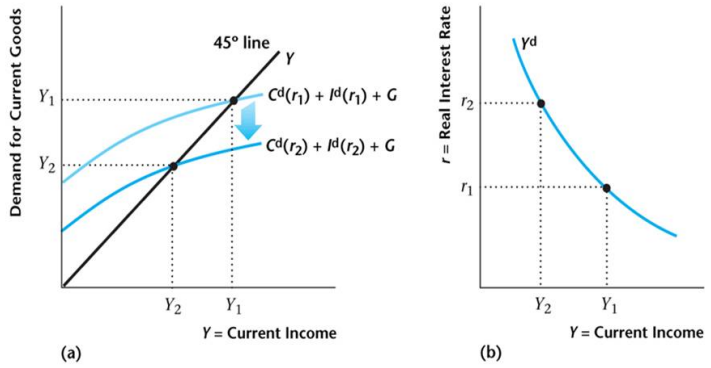


Figure 44: Construction of the Output Demand Curve, from Williamson (2011)

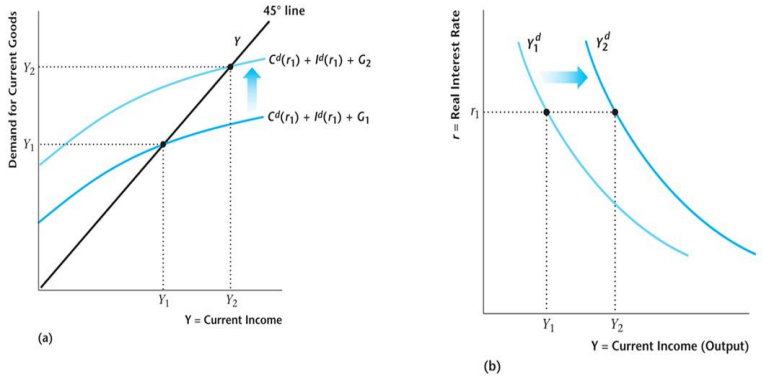


Figure 45: The Effect of an Increase in Government Spending on the Output Demand Curve, from Williamson (2011)

These feedbacks will get smaller and smaller each time and will eventually die out near the equilibrium levels represented in Figure 47. The result of an increase in future TFP (assuming the intertemporal substitution of leisure dominates income effects on leisure) is greater output, a higher real interest rate, increased employment, a lower wage rate, more investment, and an uncertain impacts on consumption.

The second experiment is to simulate an increase in credit market uncertainty. In the context of the model, this requires interpreting the representative firm as an aggregate of many different firms. This means that firms are not identical, but behave in aggregate as described by the representative firm. All of these individual firms invest, but some borrow to purchase investment goods, while others lend while still investing in future capital. The key distinction is that those firms which borrow are required to pay a risk premium above the real interest rate. The model simulates a change in credit market uncertainty by varying the risk premium, which is not explicitly modeled.

Figure 48 shows the impact of an increase in credit market uncertainty in the two-period model. Because the firms who borrow for investment purposes must pay more, their investment falls. The investment of firms who do not borrow is unchanged because the real interest rate remains the same.

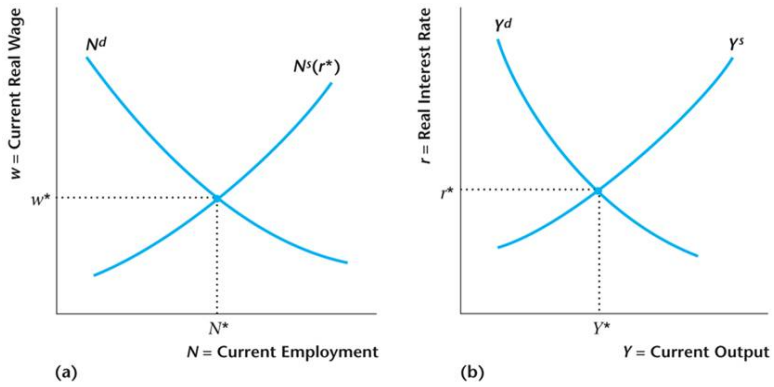


Figure 46: The Complete Two Period Model, from Williamson (2011)

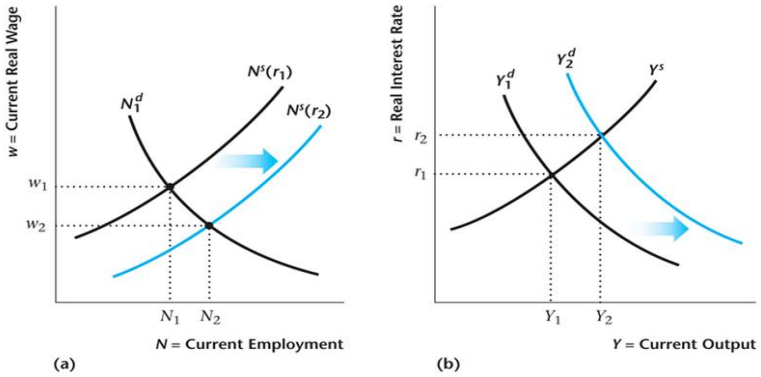


Figure 47: The Equilibrium Effects of an Increase in Future TFP, from Williamson (2011)

This leads to a shift down in the output demand curve, resulting in a lower real interest rate and output in equilibrium. As before, there are feedback effects of these new equilibrium values on the labor market. The figure shows that labor supply shifts up, as consumers choose to supply less labor. This assumes the intertemporal substitution effects of labor supply dominate the income effects (which work in the other direction). After several rounds of feedback, and assuming stronger intertemporal effects on labor supply, the result of increased credit market uncertainty is lower output, a reduced real interest rate, a higher wage rate, less employment, lower investment, and uncertain changes in consumption.

A Small Open Economy

This section extend the two-period model to the case of a small open economy. The fundamentals of the model remain the same, and the competitive equilibrium can still be summarized using graphs of the labor and current goods markets. The difference is that the real interest rate is no longer determined in the current goods market. Instead, this small open economy takes the world real interest rate (r^*) as given. This means that r is no longer the variable which adjusts to clear the

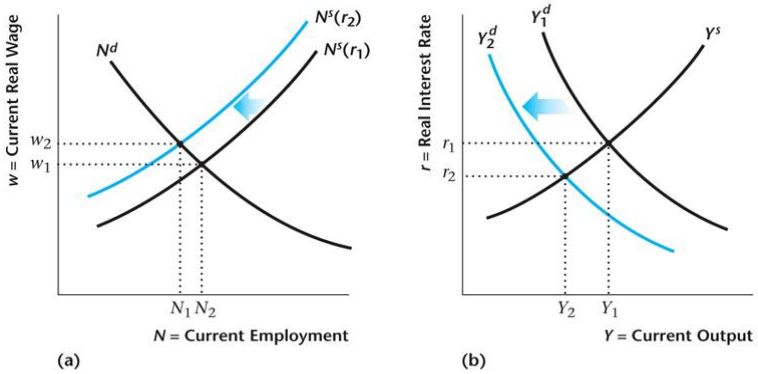


Figure 48: The Equilibrium Effects of an Increase in Credit Market Uncertainty, from Williamson (2011)

current goods market. Instead, the small economy's net exports change to equate the current supply of goods to its demand, given the world real interest rate.

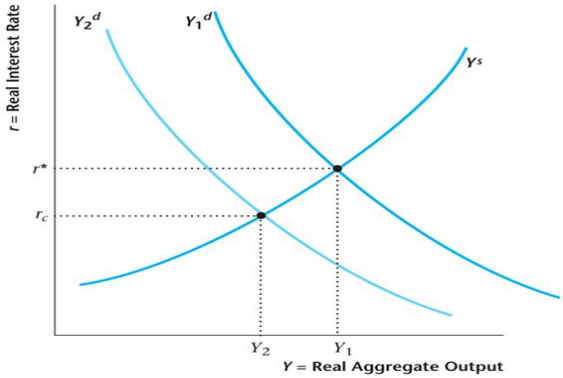


Figure 49: Equilibrium in the Current Goods Market in a Small Open Economy, from Williamson (2011)

Figure 49 shows equilibrium in the current goods market for the small open economy. While it looks similar, there are two major differences from before. The first is that r^*, the world real interest rate, is taken as given. Because of its size, the small open economy cannot change this rate. In the figure, r^c represents the real interest rate which would hold in this economy if it was closed, as in the last section. This hypothetical domestic real interest rate comes from the intersection of the output supply curve and the output demand curve in lieu of net exports (Y_1^d). That is, the export demand curve based only on domestic demand [$Y_1^d = C_1^d(Y_1^d, r) + I_1^d(r) + G_1$], as in the last section. This is sometimes referred to as absorption.

However, the actual interest rate faced by consumers and producers in this country is r^*. This means that labor supply and the demand for current goods must be consistent with the world real interest rate. Figure 49 shows that this implies an output level of Y_1 from the output supply curve. In order to achieve equilibrium, the output demand curve must shift up to a point consistent with r^* and Y_1. The

small open economy model assumes this is possible because net exports adjust to clear the current goods market.

Net exports are defined as the difference between current exports and imports, and this is a component of current demand in an open economy. That is the demand for current goods is now $Y_1^d = C_1^d(Y_1^d, r) + I_1^d(r) + G_1 + NX$, where NX represents net exports. To achieve equilibrium, NX rises to shift up the output demand curve, and the current goods market clears at the world real interest rate.

As before, there are many different experiments which can be conducted in this model. One particular experiment of interest which is not possible in the closed economy case is to vary the world real interest rate, as this is exogenous to the small open economy. The experiments are not shown here because the mechanics and interpretation are similar to those conducted above.

Dynamic Models

The two-period model gives a nice graphical representation which helps to conduct experiments and eases the interpretation of results. But this framework has little to say about the dynamics of certain variables over time. To better understand these, fully dynamic models are taken up in this section. It begins by outlining some considerations when extending beyond two periods, including choosing the length of horizon, deciding on discounting, or considerations related to expectations. The classical Ramsey model, sometimes called the neo-classical growth model, is outlined next. This is followed by an equivalent decentralized representation, which illustrates the second welfare theorem of economics. The theory is then applied through the example of a small CGE model. The final section adds uncertainty to the decentralized economy, including a characterization of the standard real business cycle model. The Solow-Swan model is derived in the appendix.

Preliminary Considerations

Extending a general equilibrium model past two periods requires making some initial decisions which will substantially influence the results. The first of these is in choosing the length of the model horizon, and this can range from a small number of periods to an infinite number. There must be a decision made on whether discounting future consumption is appropriate, and this often depends on the issue under consideration. Another choice which is always controversial is in how to expectations in the model. These can range from purely backward looking to purely forward looking.

Horizon

The horizon of a model can either be a finite number of periods or extend infinitely into the future. While both finite and infinite horizons are used in models, infinite horizons are much more popular. A finite horizon seems more natural, as consumers only live finite lives. However, when one considers

future generations it becomes difficult to decide where to stop. Should we include 2 future generations? 3? Why or why not? The fact that people generally care about their offspring means that this must be a consideration in any type of model which includes consumers.

A more pragmatic reason that infinite horizons are commonly used is simplicity. It turns out that macroeconomic models with very long horizons show similar results to infinite-horizon models. The benefit of an infinite horizon is that this representation is stationary in nature, in that the consumer and firm decisions are invariant to when decisions are made. This makes solving the model easier if appropriate mathematical tools are used. It also means that any solution should hold no matter the time period under consideration.

The distinction between a finite and infinite horizon becomes blurred when actually solving models. In CGE models the general approach is to characterize the optimality conditions of the model, calibrate the model parameters for a baseline year, and then simulate the model using the optimality conditions. The model is never really solved, but the dynamics of the variables correspond to optimal choices. As will be discussed below, a decision must be made in this case on terminal condition for the final period in the model solution. Real business cycle models are sometimes solved, in which case the finite/infinite horizon distinction is important. More often a solution is approximated using the optimality conditions. The distinction is not as important when this is done.

Discounting

Most dynamic macroeconomic models assume there is some discount on future versus current consumption. This effectively means that current consumption is more important than future consumption and the discount rate, usually denoted β quantifies this difference. This rate can also be interpreted as the weight an individual attaches to the utility of future generations in an infinite horizon problem. While the use of discounting is ubiquitous (because it seems to fit with experience and it makes model solution easier), there are some problems.

The first issue is that it is generally assumed that the discount rate does not change over time, which is unlikely to be true. People at different stages of life probably have different discount rates. The usual response to this is that while individual discount rates may vary, aggregate discount rates stay the same. There is also the question of why the consumption of future generations should have less weight than current generations. Especially when such models are used with finite natural resources, it is unclear there should be any discount at all. A final concern with such rates is in choosing their magnitude, which will have an impact on model dynamics.

Expectations

A very important component of any dynamic model which was ignored in the two-period case is how to deal with the expectations of consumers. In the two-period model consumers make decisions based

on both the current and future values of variables, for example the decision to consume or purchase bonds. Technically this is incorrect, as those decisions are made on the basis of expected future values, and these may not be known for certain. This distinction was not highlighted in the two-period model because the implicit assumption is that both consumers and firms have perfect foresight. The fact that any agents in the model know the exact values of future variables is a common assumption in CGE models, but is only possible when uncertainty is not explicitly modeled.

If uncertainty is incorporated, rational expectations are commonly used. In this case, agents in the model expect certain value for variables, and these expectations are based on all relevant information in the model. That is, agents form there expectations based on model equations, which is why rational expectations are also termed model-consistent expectations. The primary criticisms of rational expectations come from the fact that these imply that agent expectations of future values are correct on average. Although they will not always be correct due to the uncertainty in the model, any deviations from these averages are random and unpredictable. The result is that rational expectations do not differ in a systematic way from equilibrium results, so the agent's expected value of a variable is equal to the expected value predicted by the model.

Both rational expectations and perfect foresight are examples of forward-looking expectations. It is also possible to use adaptive or myopic expectations, which are backward looking. Adaptive expectations ascribe the expected value of variables to some weighted average of their past values. Myopic expectations specify that the expected value of a variable is the same as a past value of that particular variable.

The Ramsey Model

The Ramsey model is the standard dynamic general equilibrium of a closed economy. The basics of this model were first introduced in 1928 by the economist Frank Ramsey, although the version outlined here has been subsequently modified and expanded. There is no differentiation in this model between consumers, firms, or any other agents. There is a central agent (social planner), sometimes referred to as a representative agent, which is responsible for maximizing social welfare (utility). Because there is no representation of a firm, the planner acts as both a household and a firm. The idea behind this setup is to characterize the best possible allocations. What would an all-powerful planner with full information choose as the optimal allocations?

The general outline of the model is similar to the two-period one from above. The planner maximizes utility subject to a constraint. The difference is that the utility which is maximized is the discounted sum of utility infinitely far into the future, or:

$$\max_{\{C_t, N_t, I_t, K_{t+1}\}} \quad V_0 = \Sigma_{t=0}^{\infty} \beta^t U(C_t, N_t) \tag{56}$$

This states that the social planner maximizes V_0, which is the discounted utility from the current period (t=0) into the infinite future, where β is the discount rate. In order to maximize this discounted utility, the planner chooses sequences of consumption, labor to supply, investment, and future capital stock. That is, they choose how much to consume today, tomorrow, and on into the infinite future. The same is true for labor supply, investment, and future capital stock. Notice that investment and the future capital stock now belong to the planner, as there is no firm here. Both investment and capital stock in the future must be chosen by the planner because there may not be a direct correspondence between the two due to investment adjustment costs.

This type of optimization problem is common to all forward-looking dynamic general equilibrium models. Because there is no uncertainty, there is not an expectations operator outside of the summation sign, and the assumption is that the planner has perfect foresight. It is assumed here that the current time period is 0, but this problem could just as well be started at $t+35$ or $t-35$, which is due to the infinite horizon. The planner's problem is subject to the economy's resource constraint:

$$C_t + I_t + \frac{\psi}{2}\frac{I_t^2}{K_t} = F(K_t, N_t) \tag{57}$$

where

$$I_t = K_{t+1} - (1-\delta)K_t \tag{58}$$

This is an economy-wide resource constraint because the right-hand side is aggregate production, given by the production function $F(K_t, N_t)$. The left-hand side shows that this output can either be consumed or invested, with some of the investment going towards adjustment costs. The particular form chosen for investment adjustment costs assumes that these depend on the size of new investment relative to the existing capital stock. The second equation is called the capital accumulation equation. It shows that investment in the current period is the difference between capital stock in the next period and undepreciated capital stock in the current period.

This is the basic formulation of the Ramsey model. The next step is to characterize the optimal choices of the social planner by deriving first-order conditions, and then using these conditions to solve the model. Begin by formulating the Lagrangian (see the Appendix for more details):

$$L(\cdot) = \Sigma_{t=0}^{\infty}\left\{\beta^t U(C_t, N_t) + \lambda_t\left[F(K_t, N_t) - C_t - I_t - \frac{\psi}{2}\frac{I_t^2}{K_t}\right] + \mu_t\left[I_t - K_{t+1} + (1-\delta)K_t\right]\right\} \tag{59}$$

This long convoluted expression is differentiated for each time period with respect to consumption, labor supply, investment, and capital stock next period to generate the first-order conditions. There

will be three such conditions each period, one each for investment, labor supply, and capital stock next period. The first-order condition for consumption will be combined with the other three. In the Lagrangian formulation above, λ_t and μ_t are the Lagrange multipliers of each constraint, and their interpretation is that they are the change in utility given a change in the respective constraint. This means that λ_t is the marginal utility of output, while μ_t is the marginal utility of investment.

The first-order conditions with respect to investment and employment are:

$$I_t = \frac{1}{\psi}(q_t - 1)K_{t+1} \tag{60}$$

$$\frac{U_{l,t}}{U_{C,t}} = F_{N,t} \tag{61}$$

These conditions must hold for all periods at an optimum. In the first optimality condition the q_t (often called Tobin's q) is the ratio of Lagrange multipliers, $q_t = \frac{\mu_t}{\lambda_t}$.[6] The interpretation of q_t is that it is the ratio of the marginal utility of investment to the marginal utility of output. The marginal utility of investment is the benefit received by the planner of an additional unit of investment once it is installed. The marginal utility of output is the cost of this investment. If the planner produces another unit of output, this can be consumed or invested. The fact that it is used for investment (and not consumed) means that the cost is the marginal utility of the lost consumption, or λ_t. Thus q_t gives the benefit of a unit of installed capital (investment) per unit output diverted for investment. If q_t exceeds 1, then investing one unit gives a benefit greater the corresponding cost of capital, and it follows from equation (60) that investment should be increased.

The second equation is the same as in the two-period model. It states that the planner should employ labor until the marginal rate of substitution between leisure and consumption in the current period is equal to the marginal product of labor ($F_{N,t} = \frac{\partial F}{\partial N_t}$). While there is no firm in this case, it will be shown below that the marginal product of labor is equal to the wage rate for a firm at the optimum. The final condition is a messy expression due to the investment adjustment costs. Because the interpretation is similar without these costs, the simplified expression excluding investment adjustment costs is presented:

$$\frac{U_{C,t}}{U_{C,t+1}} = \beta[F_{K,t+1} + 1 - \delta] \tag{62}$$

The left-hand side of this expression is the consumer's trade-off between consuming today or

[6] In its original formulation, Tobin's q refers to the ratio of the market valuation of existing assets to the reproduction cost.

tomorrow, and is the same as the two-period case. This can be thought of as the cost of deferring consumption today. The right-hand side is the benefit from investment, the planner gets additional production equal to the marginal product of that investment in the next period ($F_{K,t+1} = \frac{\partial F}{\partial K_{t+1}}$), with some adjustment for depreciation. This equation is sometimes referred to as an Euler equation. When investment costs are included the equation is more complicated, but the fundamental interpretation stays the same.

Ideally, the first-order conditions, resource constraint, and definition of q_t could be used to solve this model for the optimal allocations of consumption, investment, employment, and output. Practically this is difficult because the model has an infinite horizon. One alternative is instead to use a finite horizon. In this case the model can be solved by assuming terminal values and then solving backwards to find the implied values of other variables. Other than the problem of choosing these terminal values, there are also concerns related to the length of horizon. The infinite horizon model could be solved using the methods of dynamic programming, but these are limited to small models.

The primary alternatives are to use the first-order conditions to trace the dynamics of model variables. These approximate solutions come in different forms. For CGE models the standard approach is to choose a base year and then to ensure the model can reproduce this base year. The model equations then trace out the variable dynamics from the base year. Real business cycle models instead use the model equations to find a steady state, and the approximate the responses of model variables as they deviate from the steady state.

The Decentralized Economy

The Ramsey problem is a good starting point for introducing dynamics, but its structure is difficult to interpret. The model remains popular, however, because the optimality conditions which characterize a solution to the Ramsey model are exactly the same as those which come from a decentralized model with consumers and firms (and no distortions). This is famous result is summarized in the second welfare theorem of economics, where the planner's allocations are the Pareto optimum which can be matched by the undistorted decentralized economy.

To see this equivalence, begin with the consumer's problem. This consumer chooses sequences of consumption, labor supply, investment, and capital next period to maximize lifetime utility:

$$\max_{\{C_t, N_t, I_t, K_{t+1}\}} \quad V_0 = \Sigma_{t=0}^{\infty} \beta^t U(C_t, N_t) \tag{63}$$

Here, it is assumed that the consumer owns the capital and rents it to firms. Because the consumer owns the firm, this is equivalent to letting the firm make investment decisions. This maximization is

subject to a budget constraint each period:

$$C_t + I_t + \frac{\psi}{2}\frac{I_t^2}{K_t} = w_t N_t + r_t K_t \tag{64}$$

where

$$I_t = K_{t+1} - (1-\delta)K_t \tag{65}$$

Notice the similarities to the planner's resource constraint. The difference is that the right-hand side of the budget constraint is the income of the consumer from supplying labor (w_t is the wage rate) and from renting capital to the firm (r_t is the rate of return on capital). This leads to identical first-order conditions with respect to investment and labor supply as in the planner's problem:

$$I_t = \frac{1}{\psi}(q_t - 1)K_{t+1} \tag{66}$$

$$\frac{U_{l,t}}{U_{C,t}} = F_{N,t} \tag{67}$$

The first-order condition for capital next period (excluding investment adjustment costs for simplicity) is slightly changed:

$$\frac{U_{C,t}}{U_{C,t+1}} = \beta[r_{t+1} + 1 - \delta] \tag{68}$$

In this equation, the future rate of return on capital replaces the marginal product of capital next period on the right-hand side, otherwise the equation is exactly the same as with the planner.

There are two steps to demonstrating equivalence between this decentralized economy and that of the planner. First, it must be shown that production is equal to income, so that the budget constraint is the same as the resource constraint. Second, the rate of return on capital must be the same as the marginal product of capital, so that the first-order conditions on capital next period are the same.

Each of these can be shown by using the firm's problem, which is the to choose labor and capital inputs to maximize profits:

$$\max_{K_t, N_t} \quad \pi_t = Y_t - r_t K_t - w_t N_t \tag{69}$$

where

$$Y_t = F(K_t, N_t) \tag{70}$$

Because the firm does not make any intertemporal decisions, this problem is the same as in the static case, but for each period. The first-order conditions are then:

$$r_t = F_{K,t}(K_t, N_t) \tag{71}$$

$$w_t = F_{N,t}(K_t, N_t) \tag{72}$$

The first equation states that the rate of return on capital is the marginal product of capital, and the second that the wage rate is the marginal product of labor. The first result allows us to show the consumer's first-order condition on capital is the same as that of the planner. It remains to show that $F(K_t, N_t) = r_t K_t + w_t N_t$ so that the budget constraint and the planner's resource constraint are the same thing. The easiest way to show this is to utilize the fact that profit is zero, and from the definition of profit it follows that production equals income.[7] The model can be solved as described for the Ramsey problem.

In summary, the previous two sections have outlined the Ramsey model and shown it is equivalent to a representative agent model that has both a consumer and firm, but no distortions. In deriving these results it was assumed the planner and/or consumer has perfect foresight over future values of variables. The q-theory of investment was also introduced as part of the derivation.

A Sample CGE Model: Finite Horizon

The sample dynamic CGE model outlined here is the same as the closed-economy model put forth in the static case, but extended over multiple periods. The optimality conditions are characterized using an infinite horizon, but the model solution is approximated by specifying initial data and allowing the model optimality conditions to drive dynamics through a finite number of periods. The addition of multiple periods only impacts the consumer's decision, as it leads to a choice between consumption and savings. As in many CGE models, the labor-leisure decision is not considered and it is assumed that the consumer supplies all of their time in labor at any wage rate. The model also has two sectors, and so there are two representative firms. The same utility and production functions are used as in the example static model.

[7] It can be shown that profit is zero by applying Euler's theorem.

The consumer's problem is to choose sequences of composite consumption and the capital stock next period to maximize discounted utility:

$$\begin{aligned} &\max_{\{\hat{C}_t, K_{t+1}\}} \quad \Sigma_{t=0}^{\infty} \beta^t \ln \hat{C}_t \\ s.t. \quad & w_t(N_{x,t} + N_{y,t}) + r_{x,t}K_{x,t} + r_{y,t}K_{y,t} = M_t = \hat{p}_t \hat{C}_t + I_{x,t} + I_{y,t} \\ & where \quad I_{x,t} = K_{x,t+1} - (1-\delta_x)K_{x,t} \\ & and \quad I_{y,t} = K_{y,t+1} - (1-\delta_y)K_{y,t} \end{aligned} \tag{73}$$

The budget constraint equates income (M_t) to expenditures. Expenditures result from purchasing the composite good at price $\hat{p}_t$ and investing in capital. Consumer income is derived from labor, $w_t N_t$ where w_t is the wage rate and N_t is labor supply to each sector, and capital income, $r_t K_t$ where r_t is the rate of return on capital and K_t is the capital stock in each sector. This capital depreciates at a fixed rate of δ in each sector each period. This set-up assumes that the consumer owns the capital stock and rents it to firms at price r_t per unit. Other than the time subscripts, the biggest difference from the static case is that the consumer can choose whether to consume or invest in the capital stock of either sector.

There are some implicit assumptions here. The first is that capital cannot move between the different sectors, as investment in one sector only builds capital in that particular sector. There are also no investment adjustment costs in this problem, which means that any investment today will be represented in the capital stock tomorrow. For this reason, investment is no longer a separate choice variable in the maximization problem of the consumer, and it can be substituted out. Also, because there is no utility from leisure, the consumer supplies all of their labor at any wage rate. This is normalized so that $N_t = 1$, or the consumer has one unit of time each period, and this is fully devoted to work. Finally, labor is mobile between sectors, which is why the wage rate equalizes and is not differentiated by sector.

As before, there is a second step to this optimization because the composite consumption good is a modeling construct. The consumer must minimize total costs of consumption by choosing the amount of good X ($C_{x,t}$) and good Y ($C_{y,t}$) to consume each period. This is subject to an "aggregator" of the two goods, where $p_{x,t}$ and $p_{y,t}$ are the prices of good X and good Y:

$$\begin{aligned} &\min_{C_{x,t}, C_{y,t}} \quad p_{x,t}C_{x,t} + p_{y,t}C_{y,t} \\ s.t. \quad & \hat{C}_t = [\gamma_c C_{x,t}^{\rho_c} + (1-\gamma_c)C_{y,t}^{\rho_c}]^{\rho_c} \end{aligned} \tag{74}$$

The difference from the static case is that this decision is made each period instead of just once. The γ_c are share parameters, which dictate how much of each good is consumed in this aggregation. The ρ_c is a parameter which represents the elasticity of substitution between the two types of goods.

Specifically, $\rho_c = \frac{(\sigma_c - 1)}{\sigma_c}$ where σ_c is the elasticity of substitution between $C_{x,t}$ and $C_{y,t}$.

Solving both the first and second stages of the consumer's problem (with a CES aggregator) leads to five optimality conditions which must be met each period, i.e. they must hold for all t:

$$C_{x,t} = \gamma_c \hat{C_t} \left(\frac{\hat{p_t}}{p_{x,t}} \right)^{\sigma_c} \tag{75}$$

$$C_{y,t} = (1 - \gamma_c) \hat{C_t} \left(\frac{\hat{p_t}}{p_{y,t}} \right)^{\sigma_c} \tag{76}$$

$$\hat{p, t} = [\gamma_c p_{x,t}^{(1-\sigma_c)} + (1 - \gamma_c) p_{y,t}^{(1-\sigma_c)}]^{\frac{1}{(1-\sigma_c)}} \tag{77}$$

$$\frac{C_t}{C_{t+1}} = \beta[r_{x,t+1} + 1 - \delta_x] \tag{78}$$

$$\frac{C_t}{C_{t+1}} = \beta[r_{y,t+1} + 1 - \delta_y] \tag{79}$$

These are the conditions which must hold for there to be any competitive equilibrium. The first two specify that demand of either good is a fraction of total consumption based on relative prices, consumption shares, and willingness to substitute. The third shows that the composite price depends on these same factors. The final two equate the marginal rate of substitution in consumption between periods (the cost of investment) to its benefit, for each sector.

Each firm's problem is exactly the same as before, it chooses capital and labor each period to maximize profit (industry X is shown here):

$$\begin{aligned} \max_{K_{x,t}, N_{x,t}} \quad & p_{x,t} Q_{x,t} - w_t N_{x,t} - r_{x,t} K_{x,t} \\ s.t. \quad & Q_{x,t} = Z_{x,t} K_{x,t}^{\alpha_x} N_{x,t}^{(1-\alpha_x)} \end{aligned} \tag{80}$$

In this problem, $Q_{x,t}$ is production of good X, $Z_{x,t}$ is total factor productivity in the production of good X, and α_x represents the capital share of output in sector X. Solving this problem for each good gives four optimality conditions each period:

$$r_{x,t} = \alpha_x \frac{Q_{x,t}}{K_{x,t}} \tag{81}$$

$$w_t = (1 - \alpha_x) \frac{Q_{x,t}}{N_{x,t}} \tag{82}$$

$$r_{y,t} = \alpha_y \frac{Q_{y,t}}{K_{y,t}} \tag{83}$$

$$w_t = (1 - \alpha_y) \frac{Q_{y,t}}{N_{y,t}} \tag{84}$$

The model also has two aggregate conditions which must be met each period. These state that use of either factor in both sectors cannot exceed the total supply of that factor, or:

$$\bar{K}_t = K_{x,t} + K_{y,t} \tag{85}$$

$$\bar{N}_t = N_{x,t} + N_{y,t} \tag{86}$$

Where $\bar{K}_t$ and $\bar{N}_t$ are the exogenously specified total labor and capital available in the economy each period. This completes the full specification of model equations. The next step is to bring these equations together using a competitive equilibrium.

Definition. *A competitive equilibrium is a sequence of values for endogenous quantities ($K_{x,t}$, $K_{y,t}$, $N_{x,t}$, $N_{y,t}$, $Q_{x,t}$, $Q_{y,t}$, $\hat{C}_t$, $C_{x,t}$, $C_{y,t}$, $I_{x,t}$, and $I_{y,t}$), values for endogenous prices ($r_{x,t}$, $r_{y,t}$, w_t, $\hat{p}_t$, $p_{x,t}$, and $p_{y,t}$), given exogenous quantities ($\bar{K}_t$, $\bar{N}_t$, $Z_{x,t}$, and $Z_{y,t}$) such that each period:*

1. *Consumers optimize;*
2. *Firms optimize;*
3. *The labor market clears;*
4. *The capital market clears;*
5. *The consumer's budget constraint is met; and*
6. *One of the goods markets clears.*

As discussed above, CGE models are usually not solved explicitly. Instead, the equations are used to characterize the dynamic behavior of variables given a base year. This base year data comes from construction of a SAM, and is exactly the same as outlined for a static CGE model. Once this base year database is constructed and all model parameters are calibrated, there are several other complications which arise due to the dynamic nature of the model.

The first of these is how to represent the future value of variables which appear in equations. A quick glance over the model equations shows that there are several equations where the future values of variables appear. Recall that the SAM only contains data for the initial year. How to choose values for these forward-looking variables that are consistent with the remainder of model equations?

There are three methods which can be used. The first is to put the model at a steady state, so that the values of all variables are constant throughout every period of the simulation. The trick here is to get the capital stock to the point where investment just equals depreciation, so the capital stock itself does not change. This is done in construction of the SAM, by making sure that $I_t = \delta K_t$ in each sector, which comes from the capital accumulation equation.

A more common approach is to let the variables in the model grow at a constant rate, say an exogenously specified GDP or population growth rate. Conceptually, this is the same as the steady state case, except that now investment and the capital stock must not stay the same, but also grow at this exogenous rate. Again, the preferred place to handle this is in the construction of the SAM, so that the initial data is already consistent with this balanced growth path.

The final method is to use different growth rates for different variables. Commonly, this means that population grows at one rate while GDP grows at another, and these exogenous rates are usually taken from outside forecasts. The best way to proceed here is to generate the initial database from the SAM consistent with one of these growth rates. The model can then be solved for this particular balanced growth path. Once this is done, the other growth rate can be imposed on the model (while keeping the first growth rate as well) and the model is solved for values consistent with each of these rates. This is the most common approach in policy applications.

Another issue that must be dealt with in the dynamic context is the terminal condition on investment and the capital stock. A natural way to proceed is to assume that investment is zero in this period and that all capital is liquidated. However, this creates very dramatic changes in model dynamics as the terminal period nears. To avoid this, many modelers specify that the rate of growth of investment must equal the rate of growth of GDP in the terminal period, or that the capital to output ratio must equal some specified number. Other conditions are used as well, with the main point being to avoid large changes in variable dynamics toward the latter periods of simulations.

A dynamic CGE model also runs into issues with respect to changing parameter values and changing technology. It is likely, particularly in longer simulations, that both the parameter values and technology representations in the model will change. There are various ways to deal with these, but none are particularly satisfactory. Finally, a common problem in these types of models is that capital adjusts very quickly to any changes in relative prices. This type of shift is not realistic because capital is largely immobile in the short-run. Many models use adjustment costs as a way to give some inertia to this adjustment process, and these were outlined in the model above.

Adding Uncertainty

Dynamic CGE models are very popular, especially in policy circles. This is because they can be made large enough to answer detailed policy-related questions. The reason their size can be expanded is because there is no representation of uncertainty. Most modelers assume perfect foresight (backward

looking expectations are also common) and solve the models using the methods described above. The academic community (and central banks) have tended to go in a different direction, and explicitly model uncertainty.

The types of uncertainty which can be modeled while still yielding a solution are restrictive. It is usually the case that any uncertainty is represented by unexpected shocks to model variables, and these shocks are normally distributed with mean zero, and their realizations are independent and identically distributed. Adding uncertainty also changes the interpretation of models. With a dynamic CGE model, any policy simulation can yield point estimates for the values of variables each period. One can change parameter values or other assumptions to get a range of these point estimates, but the main result is the point estimate. When uncertainty is added to the model, there are no longer point estimates for a given simulation. This is because each time the model is run the realizations of the shock can be different.

Instead modelers with uncertainty provide results in three ways. The first is to provide point estimates by using a few realizations of the shock. This is not common because the results can vary greatly depending on each particular realization. One can also use multiple simulations to compute the long-run properties of variables. A common exercise is to compare standard deviations of model variables to data over some time period. The final way to use these models is through impulse response analysis. The reactions and path of each variable to the shocks are plotted in these functions, and they provide some information on the importance of the shocks for each variable.

A Sample Real Business Cycle Model

The real business cycle model (RBC)is the exact same model as shown in the section on decentralized markets except that a shock to the production function is added. This seemingly small change has important implications for solution and notation of the model. The consumer's problem is similar, but they now maximize the discounted sum of expected utility (there are no investment adjustment costs here and no labor-leisure decision):

$$\max_{\{C_t, K_{t+1}\}} \quad \mathbb{E}_0 \left\{ \Sigma_{t=0}^{\infty} \beta^t \log C_t \right\} \tag{87}$$

The only difference from before is that the expectations operator is added to the summation. This maximization is subject to a budget constraint each period:

$$C_t + I_t = w_t N_t + r_t K_t \tag{88}$$

where

$$I_t = K_{t+1} - (1-\delta)K_t \tag{89}$$

One immediate question is why there is no expectations in front of future capital stock, K_{t+1}. The reason for this is that the future capital stock is decided in the current period, as it is a choice of the consumer. By next period this capital stock will be fixed.

There is only one first-order condition, and it should look familiar:

$$C_t = \beta \mathbb{E}_t \left\{ C_{t+1}[r_{t+1} + 1 - \delta] \right\} \tag{90}$$

The interpretation of this condition changes slightly, and it is shown in a slightly different manner. The idea is that the consumer should invest until the marginal utility loss of that investment (the left-hand side) equals expected marginal utility gain (the right-hand side). This is the expected marginal utility gain because there is uncertainty over future consumption and the future real interest rate. This is why the means by which the consumer forms expectations is crucial, as expectations drive the results.

The problem of the representative firm does not change, it still chooses capital and labor each period to maximize profits:

$$\max_{K_t, N_t} \quad \pi_t = Y_t - r_t K_t - w_t N_t \tag{91}$$

where

$$Y_t = Z_t K_t^{\alpha} N_t^{1-\alpha} \tag{92}$$

The difference comes in this definition of the production function. It is assumed that total factor productivity is stochastic, and its future value cannot be known with certainty. A standard stochastic process used to represent TFP is:

$$\log Z_t = \rho_z \log Z_{t-1} + \epsilon_{z,t} \tag{93}$$

The actual shock is $\epsilon_{z,t}$, which is assumed to be normally distributed with mean zero and a given variance. The shock is independent and identically distributed. In short, the consumer cannot predict the shock given any information in the model. The resulting first-order conditions are:

$$r_t = \alpha \frac{Y_t}{K_t} \tag{94}$$

$$w_t = 1 - \alpha \frac{Y_t}{N_t} \tag{95}$$

The model is brought together using a competitive equilibrium.

Definition. *A competitive equilibrium is a sequence of values for endogenous quantities (K_t, N_t, C_t, I_t, and Y_t), values for endogenous prices (r_t, and w_t), given an exogenous process (Z_t) such that each period:*

1. *Consumers optimize;*
2. *Firms optimize;*
3. *The labor market clears;*
4. *The capital market clears; and*
5. *The consumer's budget constraint is met.*

There is no drastic change in either the characterization of the model or characterization of a competitive equilibrium. The differences come when attempting to solve this model. In the CGE case, the standard approach is to use a baseline SAM for generating initial parameter and variable values, and then allowing the model equations to determine variable dynamics. The values of variables in future periods can be estimated because the model uses a perfect foresight assumption, which is no longer the case.

The standard approach with RBC models is instead to choose parameter values so that the model matches some long-run values of the data. This works by first finding the deterministic steady state values of the model. That is, the shock value is set to 1, and all other model values are assumed to be unchanging. The model is then solved for the implied steady state values of each variable. The modeler then picks the values of each parameter to match some long-run ratio. For example, it is common to choose the depreciation rate (δ) so that the long-run capital to output ratio ($\frac{K}{Y}$) is equal to 12, which is the U.S. average over the last 50 years. The modeler's discretion determines how many of the parameters to calibrate in this way. Other parameters, such as the discount rate (β) and capital share of output (α) are given standard values (0.99 and 0.36).

Once the parameter values are chosen, the problem of solving the model remains. The combination of an infinite horizon and uncertainty means that the model is difficult to solve using standard methods. The usual approach is to either use dynamic programming (for small models only) or to approximate the solution around the deterministic steady state.

Appendix

The Rational Decision Making Process

Any of the utility maximization problems outlined above is a special case of the rational decision making process of the consumer, based on the theory of consumer choice. One way to summarize this process is by assuming the consumer uses a three-step process the consumer uses when making decisions:

1. What is feasible?
2. What is desirable?
3. Choose the most desirable from the feasible.

Each of these steps are outlined in detail below, with the final step analogous to the consumer maximizing utility.

What is Feasible?

To characterize feasibility, commodities, commodity bundles, the consumption set, and the budget constraint are used. Commodities are the primitive object in consumer choice. These are the goods and services available for purchase in the market. In the simplest case, it is assumed that there are a finite number equal to L, with each commodity indexed by $l = 1, 2, ..., L$. A commodity bundle (or consumption bundle) is a list of the amounts of the different commodities, denoted by the vector $\hat{x}$. The lth entry of this vector represents the amount of commodity l consumed. For example, suppose an economy has rice, wheat, and corn as its three commodities. One possible commodity vector is $\hat{x}_1 = [1 \quad 0 \quad 5]'$. This represents one unit of rice, no units of wheat, and five units of corn. Commodities differ based on their physical characteristics, time, and location.[8]

The different consumption bundles can be grouped into a consumption set, X. This finite consumption set has as its elements the consumption bundles which the individual can consume given any physical constraints, $X = \{\hat{x}_1, \hat{x}_2, ...\hat{x}_j\}$, where j can be large but is finite. For example, an individual cannot consume a hot dog in Chicago and Paris at the same time, so a consumption bundle with this combination could not be in the consumption set. The set itself is limited by physical constraints, but the consumer also has an income constraint. This is summarized by a budget constraint, which restricts the value of purchases by the consumer to be no larger than their wealth (w). Quantities consumed are contained in the individual consumption bundles, but calculating a value requires prices. A key assumption usually made is that the consumer takes the price for good l (p_l) as given.

[8] As an example consider the same brand of hot dog made in Georgia. If one of these hot dogs is consumed on Sunday and another on Monday, they are considered different commodities. Furthermore, if one is consumed in Chicago and another in New York, they are also different commodities. These possibilities expand the number of commodities substantially.

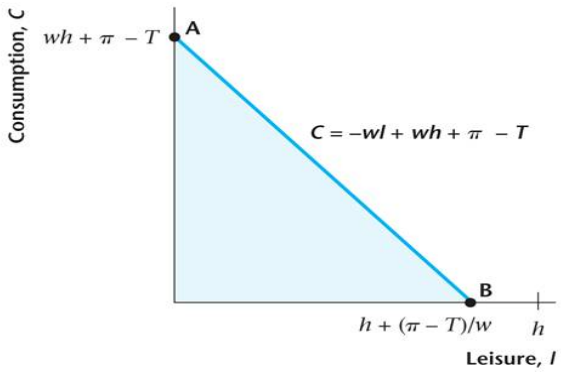


Figure 50: Representative Consumer's Budget Constraint with $T > \pi$, from Williamson (2011)

The combination of the consumption set and the budget constraint yields a new set, called the (Walrasian) budget set, B. This new set has as its elements all consumption bundles for the consumer which are feasible and affordable.[9] The prices of these goods are assumed to be given. Figure 50 shows the budget constraint from the static model outlined above, where the area under the line is the associated budget set.

What is Desirable?

How to choose the most desirable bundles in the consumption set? In the classical approach to consumer choice, one first specifies the consumer's preferences over these bundles in order to compare between the bundles. The preferences are summarized by a preference relation ($\succeq$, read "at least as good as") which is assumed to be rational.[10] Rationality of the preference relation means that it has the properties of completeness and transitivity.[11]

A consumer's preferences are complete if they are able to compare any two consumption bundles. This means they either prefer one to the other, or are indifferent between the two. This assumption is stronger than it may seem at first glance, as consumption bundles can be very different, making comparisons difficult.[12] Completeness allows the consumer to compare and rank all bundles in the consumption set. Transitivity ensures consistency in these comparisons. This assumptions states that given any three consumption bundles $\hat{x}, \hat{y}, \hat{z}$ in X, if $\hat{x}$ is preferred to $\hat{y}$ and $\hat{y}$ is preferred to $\hat{z}$, then $\hat{x}$ must be preferred to $\hat{z}$.[13] Taken together, completeness and transitivity assume that the consumer can consistently rank the bundles in their consumption set according to preferences.

[9] These are consumption bundles which already lie in X, and when associated with prices have a value less than the consumer's wealth.

[10] Technically, this is a binary relation on X.

[11] To many economists, these properties are what make a consumer "rational".

[12] Completeness: $\forall \hat{x}, \hat{y} \in X$, $\hat{x} \succeq \hat{y}$ or $\hat{y} \succeq \hat{x}$ or both.

[13] Transitivity: $\forall \hat{x}, \hat{y}, \hat{z} \in X$, if $\hat{x} \succeq \hat{y}$ and $\hat{y} \succeq \hat{z}$ then $\hat{x} \succeq \hat{z}$.

Two additional assumptions are usually made to ease computation in the third step of the process. The first is called non-satiation, which is the assumption that more is always preferred to less. This will allow the consumer to differentiate between bundles which are separated by only a small difference in commodity quantities. The other assumption is convexity of the preference relation. This amounts to stating that the consumer has a preference for diversity in their consumption bundle. It will allow a solution to the problem presented below.

The four assumptions give a complete characterization of what is desirable for the consumer. It would be much more useful, however, if the preference relation could somehow be quantified. This is done by introducing a familiar function, the utility function, which is a numerical representation of the consumer's preference relation.[14] To ensure that such a representation exists, continuity of the preference relation is also assumed. This means that the utility function will be continuous as well, and ensures there are no sudden jumps or breaks in preferences. One important note is that the utility function representing a preference relation is not necessarily unique. Any positive monotonic transformation of the utility function also represents the same preference relation. The implication is that the absolute level of utility output from any utility function is meaningless in and of itself. Interpretation requires comparison of other alternatives which come from the same utility function.

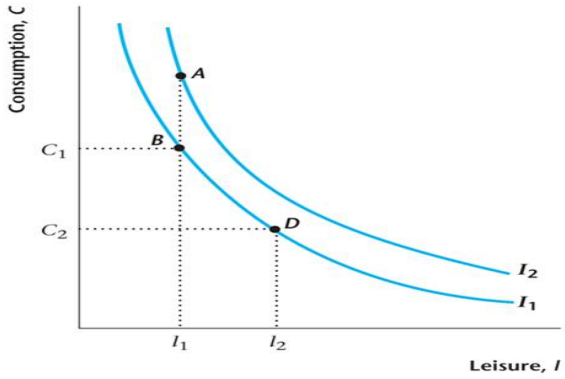


Figure 51: Indifference Curves, from Williamson (2011)

A useful way to summarize preferences graphically is by using indifference curves, as shown in Figure 51. These curves plot (in the two-good case) the different combinations of commodities which give the same level of utility. In Figure 2, all bundles of the two goods along the curve U_1 give the same level of utility. And because more is preferred to less, $I_{23} > I_1$. Given the utility function, both feasibility and desirability can now be characterized analytically (and graphically in simple cases), leading to the final step in the rational decision making process.

[14] The utility function is mapping from the consumption set to the real line, $U : X \to \Re$.

Choose the Most Desirable from the Feasible

In this final step of the rational decision making process, the consumer chooses their preferred consumption bundle from the budget set. Given the steps above, this is the same as the consumer maximizing utility (choosing the most desirable) subject to their budget constraint (from the feasible). The optimal bundle is shown graphically in Figure 52.

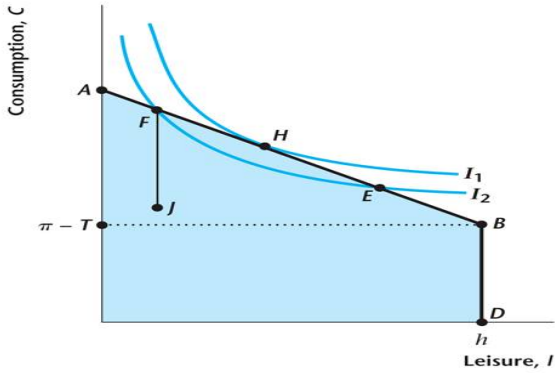


Figure 52: The Optimal Bundle, from Williamson (2011)

Constrained Optimization

This section briefly reviews the mechanics of constrained optimization using the method of Lagrange. The theory behind such optimization problems is standard, and too large to cover in this document. For illustration, the second stage of the consumer's cost minimization problem from the static CGE model is used. The problem for the consumer is to choose consumption of each particular good, C_x and C_y, to minimize the total costs of consumption. This is subject to each demand being aggregated to a composite good via a CES function. Mathematically:

$$\begin{aligned} \min_{C_x, C_y} \quad & p_x C_x + p_y C_y \\ s.t. \quad & \hat{C} = [\gamma_c C_x^{\rho_c} + (1-\gamma_c) C_y^{\rho_c}]^{\rho_c} \end{aligned} \tag{96}$$

The goal from such an optimization problem is to derive first-order conditions. These conditions describe the relationships which must hold for the consumer to be making optimal decisions. In order to arrive at these conditions, the constrained maximization problem must be solved. A simple and mechanical method to use is that of Lagrange. The first step is to define the Lagrangian, which is an equation that combines the objective function (the total costs on the first line) and the constraint (the CES aggregator):

$$L(C_x, C_y, \lambda) = p_x C_x + p_y C_y - \lambda([\gamma_c C_x^{\rho_c} + (1-\gamma_c) C_y^{\rho_c}]^{\rho_c} - \hat{C}) \tag{97}$$

The variable λ is the Lagrange multiplier. It can be interpreted as the change in the objective function given a change in the constraint. In this case it represents the marginal costs of an additional unit of composite consumption. Notice how the Lagrangian function takes the objective function and subtracts the entire constraint equation times the Lagrange multiplier. It is only a function of C_x,C_y, and λ because the prices and composite consumption are taken as given by the consumer.

The next step is to derive the first order necessary conditions. This means take the partial derivatives of the Lagrangian with respect to C_x,C_y, and λ and set these equations equal to zero. These conditions are called first-order because only the first derivative is taken. The adjective necessary means that they must hold at the optimal choice of consumption, in this case at the minimum of total costs for consumption. Each equation is set equal to zero because at either a maximum or minimum the slope of the function should equal zero, otherwise it is not a turning point.[15]

The conditions here are:

$$L_{C_x}(C_x, C_y, \lambda) = 0 = p_x - \lambda\gamma_c C_x^{\rho_c - 1}[\gamma_c C_x^{\rho_c} + (1-\gamma_c)C_y^{\rho_c}]^{\rho_c - 1} \tag{98}$$

$$L_{C_y}(C_x, C_y, \lambda) = 0 = p_y - \lambda(1-\gamma_c) C_y^{\rho_c - 1}[\gamma_c C_x^{\rho_c} + (1-\gamma_c)C_y^{\rho_c}]^{\rho_c - 1} \tag{99}$$

$$L_{\lambda}(C_x, C_y, \lambda) = 0 = [\gamma_c C_x^{\rho_c} + (1-\gamma_c)C_y^{\rho_c}]^{\rho_c} - \hat{C} \tag{100}$$

These complicated expressions can be rearranged to solve for the three unknowns (in terms of prices and composite consumption). The algebra is conceptually straightforward but messy, so is not shown here. After some simplification these yield the three optimality conditions from the small CGE model in the text:

$$C_x = \gamma_c \hat{C} \left(\frac{\hat{p}}{p_x}\right)^{\sigma_c} \tag{101}$$

$$C_y = (1-\gamma_c) \hat{C} \left(\frac{\hat{p}}{p_y}\right)^{\sigma_c} \tag{102}$$

$$\hat{p} = [\gamma_c p_x^{(1-\sigma_c)} + (1-\gamma_c) p_y^{(1-\sigma_c)}]^{\frac{1}{(1-\sigma_c)}} \tag{103}$$

The same procedure can be used to solve most optimization problems found in general equilibrium

[15] There are several assumptions and conditions which must be met for this maximum or minimum to exist which are not discussed here. These are covered in any basic textbook on mathematical economics, such as Sydsaeter and Hammond (2008).

models, including those which are dynamic.

The Elasticity of Substitution

The elasticity of substitution is arguably the most important parameter used in applied general equilibrium models. When used in the context of consumption, it is roughly the percentage change in demand between two commodities given a 1% change in the marginal rate of substitution between the goods. The definition is similar when used in production. Here, it is the percentage change in demand between two inputs given a 1% change in the marginal rate of transformation between the inputs.

The formula for the elasticity of substitution between two goods in consumption (say consumption and leisure) is:

$$\sigma_c = \frac{\partial \log\left(\frac{l}{C}\right)}{\partial \log MRS_{lc}} \tag{104}$$

The formula for the elasticity of substitution between two inputs in production (say labor and capital) is similar:

$$\sigma_p = \frac{\partial \log\left(\frac{N}{K}\right)}{\partial \log MRTS_{KN}} \tag{105}$$

As with any elasticity, there are no units attached to the elasticity of substitution. A good way to interpret the formula is to use the fact that the log difference of small quantities approximates the percentage change in those quantities. Then the denominator for the elasticity of substitution in consumption is the percent change in the willingness of the consumer to substitute between leisure and consumption. The denominator for the elasticity of substitution in production is the percent change in the technological ability of the firm to substitute between capital and labor.

Suppose the MRS_{lc} rises by one percent. This means that consumers are more willing to give up consumption goods for one unit of leisure, or the relative price of consumption is higher. Recall from the static model above that the MRS_{lc} is equal to the wage rate. The value of the elasticity of substitution will depend on the response of demand for l relative to C to this change in the wage rate. If this demand also rises by 1%, then the elasticity is equal to 1, called unit elasticity. This result says that a 1% rise in the wage rate, which means that the price of consumption is 1% higher, leads to a 1% increase in the demand for leisure relative to consumption. If the demand for l relative to C rises by 2.5%, then the elasticity is 2.5, which is a common value used for the elasticity of substitution between two goods in consumption.

The production case is similar. Suppose the $MRTS_{KN}$ rises by one percent. This means that firms are more willing to give up labor for one unit of capital. One can think of this as a rise in the relative

price of capital. This is because capital is relatively scarce compared to labor and has a higher marginal productivity (hence the higher relative price). Firms are able to use this relatively smaller amount of capital but still produce the same amount of output. If the demand for labor relative to capital rises by 1% in response, this is the unit elastic case, which corresponds to the Cobb-Douglas production function.

Elasticities of substitution in production between inputs lower than one are commonly used. These inputs are called "weak substitutes", which implies that there is little change in the relative demand for the inputs given large changes in the $MRTS_{KN}$. As the elasticity of substitution goes to zero, the inputs become perfect complements. This means that the two goods must be used together. A production function with an elasticity of substitution of zero between its inputs is called a Leontief function. This type of function is often used as the top level of a production structure in CGE models, where there is no substitution allowed between factors of production and intermediate goods.

To show that the Cobb-Douglas function has an elasticity of one, define this as: $F(K,N) = ZK^{\alpha}N^{1-\alpha}$. The marginal rate of transformation of capital for labor, $MRTS_{KN}$, is the slope of the isoquant of this function. The isoquant is the production analogue to an indifference curve. Bundles on an isoquant show all of the capital/labor combinations which yield the same level of production. As the marginal rate of transformation is the slope of the indifference curve, the slope of the isoquant is the marginal rate of transformation.

To find this slope the function $F(K,N)$ can be implicitly differentiated. Along the isoquant we know that output is constant, or $F(K,N) = c$. Then by application of the chain rule, the total derivative of this equation can be written as:

$$dc = \frac{\partial F(K,N)}{\partial K}dK + \frac{\partial F(K,N)}{\partial N}dN \tag{106}$$

Because output is constant along an isoquant, $dc = 0$. Using this, the expression can be rearranged to give the slope along the isoquant:

$$\frac{dN}{dK} = -\frac{\frac{\partial F(K,N)}{\partial N}}{\frac{\partial F(K,N)}{\partial K}} \tag{107}$$

The right-hand side is obtained by differentiating the production function, so that:

$$\frac{\partial F(K,N)}{\partial N} = (1-\alpha)\frac{F(K,N)}{N} \tag{108}$$

$$\frac{\partial F(K,N)}{\partial K} = \alpha\frac{F(K,N)}{K} \tag{109}$$

Substituting this back into equation (107) yields the $MRTS_{KN}$:

$$MRTS_{KN} = \frac{dN}{dK} = -\frac{1-\alpha}{\alpha}\frac{K}{N} \tag{110}$$

This expression can be substituted into the definition for the elasticity of substitution:

$$\sigma_p = \frac{\partial \log \frac{N}{K}}{\partial \log MRTS_{KN}} = \frac{\partial \log \frac{N}{K}}{\partial \log (\frac{1-\alpha}{\alpha}\frac{N}{K})} \tag{111}$$

The right-hand side is equal to one because the constant can be factored out and does not change. This leaves the ratio of the change in the log of the demand for labor relative to capital, which is 1. The denominator shows the fraction $\frac{N}{K}$ instead of $\frac{K}{N}$ as in equation (110) because the negative sign has been taken through.

The Solow-Swan Model

The Solow-Swan model is the standard model of economic growth. It is similar to the dynamic general equilibrium models described in the main text, save that there is no explicit optimization by consumers or firms. In fact, there is also no explicit differentiation between consumers and firms. The Solow-Swan model assumes there are many consumers and that an economy's production is characterized by an aggregate production function, which is equivalent to aggregate income. The model is dynamic, but the steady state can be shown graphically, and this representation is sufficient to understand its important points. The Solow-Swan model helps to shape the long-run properties of many macroeconometric models used widely for policy analysis.

In what follows the key equations of the model are set out and manipulated to yield the primary equation which summarizes model dynamics. In the steady state, this equation is then shown graphically, and some simple experiments are conducted. The Solow-Swan model assumes a constant and exogenous population growth rate:

$$N_{t+1} = (1+n)N_t \tag{112}$$

N is the population, and it grows at the constant rate of n each period. This equation can be re-written as: $\frac{N_{t+1}}{N_t} = 1+n$. In aggregate all of the consumers in the economy can either consume (C_t) or save (S_t) aggregate income (Y_t):

$$Y_t = C_t + S_t \tag{113}$$

Because this is a closed economy, savings must equal investment by definition $(S_t = I_t)$, which will be

used below. Importantly, it is also assumed that consumers as a whole save a fraction of aggregate income, or:

$$C_t = (1 - s)Y_t \tag{114}$$

In this equation s is a fixed savings rate for the aggregate economy. This is where the Solow-Swan model diverges from general equilibrium models: there is no endogenous savings/consumption decision. Because of this, it is not necessary to model the preferences of the consumers, which is why the model is solved in aggregate terms. The fixed savings rate implies that savings is a fixed fraction of aggregate income:

$$S_t = sY_t \tag{115}$$

This completes the description of the consumers in the model. There is no explicit representation of the firm, but the model does have an aggregate production function:

$$Y_t = Z_t F(K_t, N_t) \tag{116}$$

As is standard, production depends on capital and labor, and can be altered by technological progress. Capital also follows the standard accumulation equation:

$$K_{t+1} = K_t(1 - \delta) + I_t \tag{117}$$

Capital in the next period is investment plus any undepreciated capital. At this point, achieving consistency in the model requires assuming that all markets clear. There are two markets, one for consumption of current goods and a capital market (there is also a labor market, but this clear by assumption in the model). The market clearing condition for the capital market was stated above, that savings equals investment. In the current goods market, the clearing condition states that any production must be used for consumption or investment:

$$Y_t = C_t + I_t \tag{118}$$

This is the equation which can be manipulated to generate diagrams. First, substitute out both consumption and investment using two of the equations above to give:

$$Y_t = (1 - s)Y_t + K_{t+1} - K_t(1 - \delta) \tag{119}$$

Next, substitute out income and insert the production function:

$$Y_t = (1-s)Z_t F(K_t, N_t) + K_{t+1} - K_t(1-\delta) \tag{120}$$

This expression can be rearranged to put it in terms of capital next period:

$$K_{t+1} = sZ_t F(K_t, N_t) + K_t(1-\delta) \tag{121}$$

There is one more step which is standard before obtaining a graphical representation, which is to put the variables in per-capita terms. This is done because the model does not have an explicit representation of welfare, and income per capita is used as a proxy. To put in this form, divide both sides by N_t:

$$\frac{K_{t+1}}{N_t} = \frac{sZ_t F(K_t, N_t) + K_t(1-\delta)}{N_t} \tag{122}$$

Notice that the left-hand side has $t+1$ over t, which gives inconsistent dimensions. To get around this, multiply the left-hand side by 1, which is the same as $\frac{N_{t+1}}{N_{t+1}}$:

$$\frac{K_{t+1}}{N_{t+1}}\frac{N_{t+1}}{N_t} = \frac{sZ_t F(K_t, N_t) + K_t(1-\delta)}{N_t} \tag{123}$$

Now use the fact that $\frac{N_{t+1}}{N_t} = 1+n$ and substitute in the left-hand side. Also, because the production function has constant returns to scale, the N_t can be factored out of this equation when in per-capita terms. The final equation, with lower-case reflecting per-capita variables is:

$$k_{t+1} = \frac{sZ_t f(k_t) + k_t(1-\delta)}{(1+n)} \tag{124}$$

This is the key equation in the Solow-Swan model. It describes the evolution of capital per person in the economy. Capital per-capita is the most important variable, because as shown in the equation, it determines output per-capita. So higher capital per person means a higher standard of living in the model. A good place to begin analysis is to plot capital between two adjacent periods and derive the steady state level of capital, as in Figure 53.

The steady state level of capital stock is that which the economy tends to in the long-run. In this model, if the current capital stock is below the steady state level (k^*), investment exceeds capital depreciation so that the capital stock grows. To see this on the diagram, pick any point k_t below k^*. Project this point onto the figure, and the value of capital next period is above the 45 degree line. Because the 45 degree line gives all of the points where $k_t = k_{t+1}$, any points above this line indicate

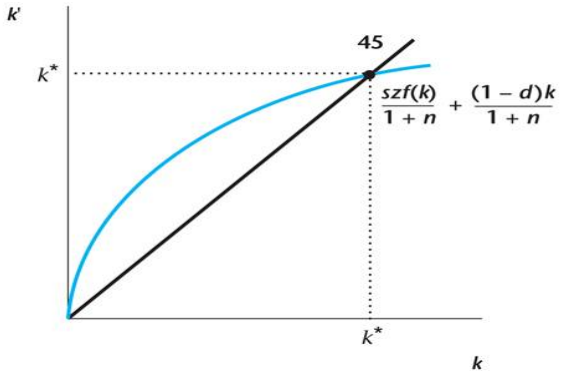


Figure 53: Steady State Capital in the Closed Economy Solow-Swan Model, from Williamson (2011)

the capital stock is large in the next period.

This graph can be used to experiment with the model by changing various inputs and assessing their impacts on the steady state level of capital stock. However, while this representation makes the steady state level of capital clear, it is not the best way experiment with the model. For a better representation, use equation (124) but assume the model is at steady state, so that $k_t = k_{t+1} = k^*$. This gives:

$$szf(k^*) = (n + d)k^* \tag{125}$$

At the steady state this relationship must hold, so that per-capita investment (the left-hand side) is equal to the steady state level of capital when population growth and depreciation are accounted for. This relationship can be plotted, with steady state level of capital on the horizontal and either the left or right-hand side expressions on the vertical (by definition they are equal). This is shown in Figure 54.

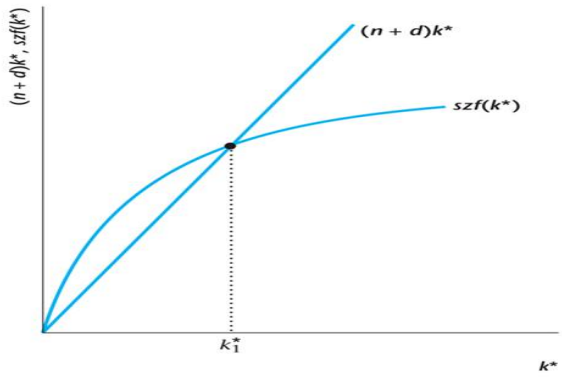


Figure 54: An Alternative Representation of Steady State Capital in the Closed Economy Solow-Swan Model, from Williamson (2011)

The shape of the production function gives the left-hand side its shape, while the right-hand side is linear as represented in the figure. We are now in a position to change any of the exogenous variables (n, d, s, orz) to see the impact they have on the steady-state level of capital, and therefore long-run living standards. Figure 55 shows the impact of an increase in the savings rate.

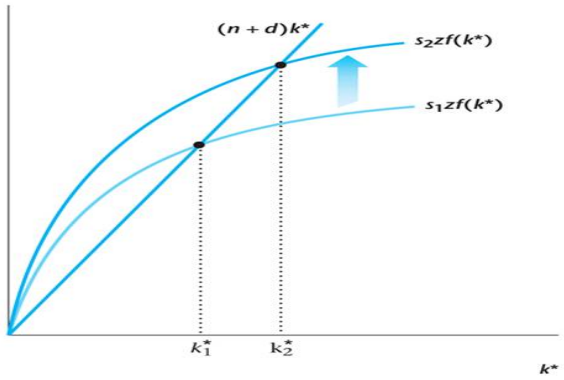


Figure 55: An Increase in the Savings Rate in the Closed Economy Solow-Swan Model, from Williamson (2011)

This increase will shift up the curve, and leads to a higher level of capital per person, which leads to higher GDP per person in the long-run. Thus increased savings should increase standards of living. This happens because consumers give up some consumption today to build capital, which has a higher payout in the future. While this is an intuitive result, there is only so high the savings rate can rise, so there are limits to this type of growth. The next experiment increases either the population growth rate or the depreciation rate of capital (or both) as in Figure 56.

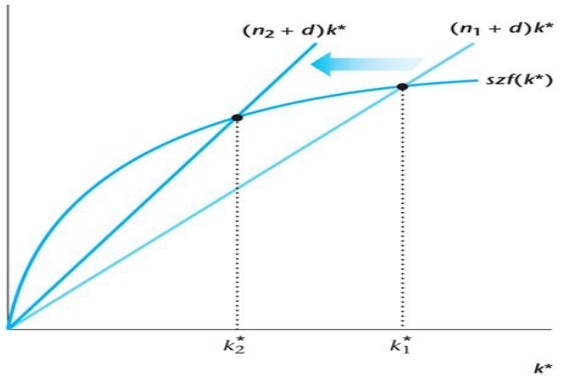


Figure 56: An Increase in the Population Growth Rate or Depreciation Rate in the Closed Economy Solow-Swan Model, from Williamson (2011)

This shifts up the other curve and results in a lower level of capital in the long-run. This result is also intuitive, in that if there are more people or capital wears out faster, there is less capital to go around per person. This is what is reflected in the graph. The final experiment gives the most important

lesson from the Solow model, as it increases TFP as in Figure 57.

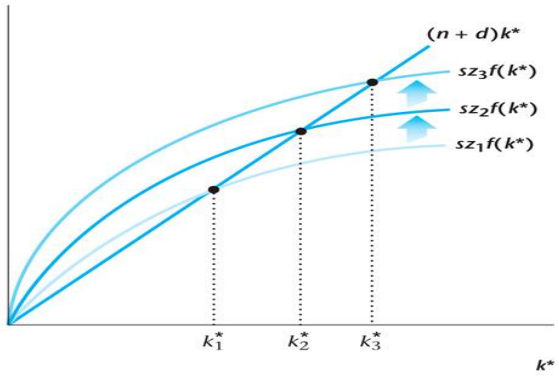


Figure 57: An Increase in Total Factor Productivity in the Closed Economy Solow-Swan Model, from Williamson (2011)

Increases in total factor productivity continually increase living standards through more capital per person. The strength of this type of growth is that it is potentially unlimited.

References

Shoven, John B. and John Whalley, *Applying General Equilibrium*, 1st ed., Cambridge University Press, 1992.

Sydsaeter, Knut and Peter Hammond, *Essential Mathematics for Economic Analysis*, 3rd ed., Prentice Hall, 2008.

Wickens, Michael, *Macroeconomic Theory: A Dynamic General Equilibrium Approach*, 1st ed., Princeton University Press, 2008.

Williamson, Stephen D., *Macroeconomics*, 4th ed., Pearson, 2011.

Appendix B: The Mechanics of Vector Autoregressions

Overview

This document summarizes some basic technical details of vector autoregressions (VARs) for use in both policy analysis and forecasting.[1] To make the document as self-contained as possible, the first section briefly provides a short introduction to basic time series topics, univariate time series models ($ARMA$ models), estimation, Bayesian methods, and forecasting. The second section gives a general description of a VAR, outlines its estimation, and discusses alternative representations of the VAR process. The following section takes up basic policy analysis using VARs, with a focus on identification. The strengths and weaknesses of using this approach for policy analysis are also covered, as is identification using Bayesian techniques. The fourth section outlines forecasting with VARs, including some pros and cons, as well as additional detail on forecasting using Bayesian methods. The final section reviews alternative techniques for identification in VARs when used for policy analysis. [2]

Preliminaries

This section briefly reviews some time series concepts that are used in the context of vector autoregressions (VARs). The discussion below selectively follows Kennedy (2008) and Enders (2010), but discussion of the concepts can be found in any textbook on time series analysis. Cochrane (2005) provides a more advanced treatment with derivations of key concepts.

Basic Time Series Concepts

VARs are based on stochastic difference equations. A difference equation expresses the value of a variable as a function of its own lagged values, other variables, and time. The equation becomes stochastic if any of the other variables are random, or if random error or disturbance terms are added. The implications of the efficient market hypothesis for stock prices are consistent with a well-known

[1] None of the material covered here is original. As much as possible, the original sources of the equations, explanations, and examples have been cited.

[2] For more in depth treatment, the reader is referred to an advanced textbook on time series analysis such as Lutkepohl (2007) or Enders (2010).

example of a stochastic difference equation, the random walk:

$$y_t = y_{t-1} + \epsilon_t \tag{1}$$

In this case y_t is interpreted as the price of a share on day t and ϵ_t as a random disturbance term with mean zero.

The stochastic error, often called a disturbance or innovation, is very important in time series econometrics, particularly in autoregression analysis. The properties of this process will often be restricted so that it is white noise. A white noise process has zero mean, finite variance, and is serially uncorrelated. The fact that it has zero mean says that over repeated samples the average value of the process will be zero.[3] Looking at equation (1), this means that the stock price today on average is the same as the stock price yesterday.

Recall that the variance of a process measures how far on average it deviates from the mean. It is the expected value of the squared deviation of a sample draw from its mean.[4] The variance is often assumed to be constant over time, or homoscedastic. Serial correlation, also known as autocorrelation, is the correlation between values of the process at different points in time. Correlation, which is based on covariance, measures how much two random variables move together, and ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).[5]

A common additional restriction is to assume that the realizations of the white noise process are distributed normally, making it a Gaussian white noise process. The purpose of assuming that error terms are restricted to be Gaussian white noise is to ease estimation by ordinary least squares (OLS) or maximum likelihood (ML). Not only can the estimation techniques be applied in a straightforward manner, but the resulting standard errors can be used for hypothesis testing (assuming the other assumptions of either technique are met). White noise is also a convenient assumption because such a process is stationary. A particular variable y_t is said to be stationary if the sequence of its realizations $\{y_t\}$ yields a constant mean, variance, and autocovariance. Additional realizations of the process do not change its expected value, the spread of the realizations from the mean stays the same, as does the correlation between different realizations at any point in time.

While it may be plausible to restrict error terms to be stationary, most macroeconomic aggregates of interest are not stationary. GDP, consumption, and other important macroeconomic variables grow

[3]The expected value of a process of discrete random variables $\{X_t\}$ is $E[X] = x_1p_1 + x_2p_2 + ... + x_kp_k$, where the x_j are any values that X_t can take and the p_i are their associated probabilities.

[4]Let $E[X] = \mu$, then the variance of $\{X_t\}$ at any point in time t is $Var[X_t] = E[(X_t - \mu_t)^2]$. This is often denoted σ^2 if it is constant over time, or σ_t^2 if it varies. The standard deviation, σ is the square root of the variance.

[5]The autocovariance of $\{X_t\}$ over one period is given by $Cov[X_t, X_{t-1}] = E[(X_t - \mu_t)(X_{t-1} - \mu_{t-1})]$. The autocorrelation is a normalization of this number to lie between -1 and 1, and is computed by dividing the autocovariance by $\sigma_t\sigma_{t-1}$, or the product of the standard deviations of the process at the different times.

over time -they have a trend. Trends are differentiated between those which are deterministic and those which are stochastic. Separating an arbitrary stochastic difference equation into three parts is a good way to see this clearly:

$$y_t = \text{trend} + \text{stationary component} + \text{noise} \tag{2}$$

If a series is stationary, then there is no trend component and the noise is also stationary. However, if the series is trending (and the noise component is assumed to be stationary), then the trend component can either be deterministic or stochastic. If it is deterministic, the trend can be predicted, and any deviations from the trend will be relatively short-lived. The basic idea is that the impact of noise on the trend dies out over time, so that the realizations of the time series return to the deterministic trend (plus the stationary component). This is called a trend-stationary process. If some long-run growth rate of real GDP exists in the economy, then this growth rate is a trend-stationary process.

A stochastic trend cannot be predicted, and any deviations from the trend may or may not be short-lived. The reason for this is that the impact of the noise on the trend has a permanent effect, and the noise is random by construction, making the trend random as well. An example of a model with a stochastic trend is the random walk model from above. The implication for stock prices is that the history (i.e. their past trend) is likely a poor guide for the future. A seminal paper by Nelson and Plosser (1982) showed that most macroeconomic aggregates of interest are not trend-stationary, but appear to have a stochastic trend. The growth rates of these aggregates, however, may be stationary.

In general, a non-stationary series can be made stationary by differencing. If the series needs to be differenced once to be made stationary (first differenced) it is called integrated of order one or I(1), if it needs to be differenced twice it is I(2), and so forth. A stationary series is I(0). A trend-stationary series can be made stationary without differencing by removing the deterministic trend, although finding this trend may not be straightforward. As an example, some series have time trends in the sense that they grow by some deterministic amount over time. If the trend can be found, removing it from the data is relatively straightforward.

Care must be taken to remove a trend in the appropriate manner. If a deterministic trend is removed from a model with a stochastic trend, the series is still not stationary. Trying to difference a trend-stationary series causes problems as well. In this case, another non-stationary process may be introduced into the series. A third approach to removing a trend is to use a filter designed to separate trend from cycle. Popular filters include the Hodrick-Prescott (HP) filter and various Band-Pass (BP) filters (Enders, 2010).

An easy way to look for a unit root (non-stationarity) in a time series process is to inspect the autocorrelation function. This is often called a correlogram, and is a plot of the autocorrelations of a

process over time. That is, it plots $corr(y_t, y_{t-1})$,$corr(y_t, y_{t-2})$, ..., $corr(y_t, y_{t-n})$. If the series is stationary, this should go to zero quickly, meaning that past values of y do not have an important impact on the current values. A slow decay in the correlogram over time is a good indication that the process is not stationary.

While differencing an I(1) process can make it stationary, this can entail some costs in a multivariate setting. In particular, some pairs of non-stationary variables tend not to drift too far apart because there are forces that keep them together. When such variables are individually I(1) but a linear combination of them is I(0), they are termed cointegrated. Think of consumption and disposable income, imports and exports, and many other important macroeconomic relationships. Differencing each of these variables individually can throw away important information on this relationship. A way around differencing non-stationary data can be found in this case by using an error-correction model (ECM), which is outlined below.

The concept of Granger causality is often used in the context of VARs. Roughly, one can say that a variable Granger causes another if an unexpected movement in the first variable helps to forecast the other variable. This is not causality in the usual sense because there might be other variables which affect both of those being tested. There are several ways to test for Granger causality, and these are outlined in Cochrane (2005).

It is also useful to differentiate between structural and reduced-form equations. A structural equation is one which expresses an endogenous variable y_t as being dependent on the current realization of another endogenous variable x_t, and possibly its own lags, the lags of other endogenous variables, current and past values of exogenous variables, and disturbance terms. A reduced-form equation is similar, but any included endogenous variables must be lagged. A reduced form equation does not express one endogenous variable in terms of the current value of another endogenous variable.

Summarizing the variance of a vector autoregressive process requires introduction of the variance-covariance matrix. A variance-covariance matrix, usually denoted $\mathbf{\Sigma}$, contains the covariances between different elements from a vector of random variables. For example, consider a vector of white noise processes:

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

The variance-covariance matrix for this random vector has as its (i, j) element the covariance between the ith and jth elements of the vector. So the $(1, 2)$ element of the variance-covariance matrix for the white noise processes above is $cov(\epsilon_1, \epsilon_2)$. This can be written out fully:

$$\mathbf{\Sigma} = \begin{bmatrix} cov(\epsilon_1, \epsilon_1) & cov(\epsilon_1, \epsilon_2) \\ cov(\epsilon_2, \epsilon_1) & cov(\epsilon_2, \epsilon_2) \end{bmatrix} \equiv \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

The diagonal elements are the variances of the respective elements of the random vector, which are constant over time because of the white noise assumption.[6] Similarly, the off-diagonal elements are the covariances, which are zero because white noise processes are uncorrelated by definition. Variance-covariance matrices of estimators are frequently used, particularly for OLS.

Finally, a convenient way to cut down on notation in time series analysis is to use lag operators. This operator moves a variable back the specified number of periods, i.e.:

$$L^i y_t = y_{t-i} \tag{3}$$

Putting the lag operator, L^i, before a variable y_t will lag that variable by i periods. A negative lag means it will move the variable forward by i periods. As another example, the random walk model from above can be written using a lag operator:

$$y_t - Ly_t = \epsilon_t \quad \equiv \quad A(L)y_t = \epsilon_t \tag{4}$$

where $A(L) = (1 - L)$.

ARMA Models

Autoregressive moving average (ARMA) models are very important in time series analysis, and form the basis for vector autoregressions. They are a combination of both autoregressive processes and moving average processes. An autoregressive process with p lags [$AR(p)$] is one where a variable (y_t) is dependent on only its own p lags and a disturbance term. For example, an AR(2) process is written:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + c_0 \epsilon_t \tag{5}$$

where a_0, a_1, a_2, and c_0 are unknown coefficients to be estimated and the ϵ_t are Gaussian white noise. A moving average process with q lags [$MA(q)$] is written only in terms of the disturbances, with q lags. For example, an MA(2) is written:

$$y_t = \sum_{i=0}^{2} c_i \epsilon_{t-i} \tag{6}$$

[6] The covariance of a variable with itself is the same as the variance.

An ARMA process is a combination of autoregressive and moving average processes. An $ARMA(p,q)$ is a combination of an $AR(p)$ and $MA(q)$:

$$y_t = a_i y_{t-1} + ... + a_p y_{t-p} + c_i \epsilon_{t-1} + ... + c_q \epsilon_{t-q} \tag{7}$$

Alternatively, the individual autoregressive or moving average process are special cases of the ARMA process. In this case one can think of an $MA(q)$ as an $ARMA(0,q)$, and an $AR(p)$ as an $ARMA(p,0)$.

Vector autoregressions often use the fact that stationary $AR(1)$ processes can be represented as an $MA(\infty)$. That is, there is a relationship between autoregressive and moving average processes. This can be shown by starting with an arbitrary $AR(1)$ process at t:

$$y_t = a_i y_{t-1} + c_i \epsilon_t \tag{8}$$

This same process is assumed to hold over all periods, so it can be moved back to get an equation for y_{t-1}:

$$y_{t-1} = a_{i-1} y_{t-2} + c_{i-1} \epsilon_{t-1} \tag{9}$$

This can be used to substitute equation (9) into equation (8):

$$y_t = a_i(a_{i-1} y_{t-2} + c_i \epsilon_t) + c_{i-1} \epsilon_{t-1} \tag{10}$$

The procedure is continually repeated for y_{t-2}, y_{t-3}, ..., y_{t-n}. Notice how each successive substitution adds an additional error term to the equation, moving it closer to a moving average representation. These substitutions also add the product of the coefficients on the lagged variables ($a_i a_{i-1}$ in equation (10) above). In order to write this $AR(1)$ as a moving average process, this product must go to zero as the series goes further and further back into the past. Another way to say this is that as $n \to \infty$, $a_i, a_{i-1}, ..., a_{i-n} \to 0$.

The reason that only a stationary $AR(1)$ process has an equivalent $MA(\infty)$ representation is that the product of coefficients only goes to zero if the coefficients are less than one. This is a necessary condition for the process to be stationary. The interpretation of the coefficient being less than one is that the impact of past y values decreases over time. That is, lagged values of y do not have permanent impacts on the value of y far into the future. Taking the limit of the successive iterations

from above gives:

$$y_t = \sum_{i=0}^{\infty} c_i \epsilon_{t-i} \tag{11}$$

The interpretation of the coefficient values from this representation is important when using vector autoregressions. Each c_i summarizes the impact of a one unit movement in ϵ_{t-i} on the current value of y_t. For example, c_0 gives the impact of a one unit movement in the current disturbance term on the current value of y_t. This is often called the instantaneous impact.

This process can also be written using the lag operator:

$$y_t = B(L)\epsilon_t \tag{12}$$

where $B(L) = c_0 + c_1 L + c_2 L^2 + ...$

Estimation

The primary method of estimating VARs is ordinary least squares. Under some general assumptions this is the same as maximum likelihood estimation. Each of these is briefly reviewed below, along with state space models, as are some desirable properties of estimators. Each of the estimation techniques is used in the exposition on VARs and the criteria for estimators reviews the concept of a sampling distribution, which is important in understanding the difference between classical and Bayesian estimation.

Ordinary Least Squares

In most cases, the coefficients in a VAR are estimated by ordinary least squares, just as with standard regressions. Recall that in the time series context, a general OLS equation with n observations and two independent variables is written as:

$$y_t = B_0 + B_1 x_t + B_2 z_t + u_t \quad t = 1, 2, ..., n \tag{13}$$

Here, B_0 is a constant, B_1 is a coefficient which gives the impact of the independent variable x_t on the dependent variable y_t, B_2 is also a coefficient with a similar interpretation for z_t, and u_t is the error term.

This equation can also be written in vector form:

$$y_t = \hat{J}_t \hat{B} + u_t \quad t = 1, 2, ..., n \tag{14}$$

where $\hat{J}_t = (1, x_t, z_t)$ is a 1x3 vector of independent variables, and $\hat{B} = (B_0, B_1, B_2)'$ is a 3x1 vector of coefficients. In what follows, a hat indicates a vector and bold type a matrix. The notation can be consolidated further in matrix form if the n observations are included:

$$\hat{y} = \mathbf{J}\hat{B} + \hat{u} \tag{15}$$

Now $\hat{y}$ is the nx1 vector of observations of the dependent variable, $\hat{B}$ is still the 3x1 vector of coefficients, and $\hat{u}$ is the nx1 vector of unobserved errors. The nx3 matrix **J** contains the observations of independent variables, where each row corresponds to one observation and the columns are the independent variables. The method of ordinary least squares provides estimates of the constant and coefficient values by minimizing the sum of squared errors from the system above.

There are five basic assumptions made when using OLS, which must also hold when estimating VARs. First, it is assumed that the dependent variable can be calculated as a linear function of the independent variables plus a disturbance term. For VARs, this assumption is slightly modified in that the endogenous variables are assumed to be linear functions of the other endogenous variables. Second, the expected value of the disturbance term is zero. Third, the disturbance terms have constant variance and are uncorrelated with each other.

Fourth, the observations of the independent variables can be considered fixed in repeat samples. This assumption is often weakened so that it is required only that the independent variables (or variables on the right-hand side of the regression) are uncorrelated with the error term. This point will be important for VAR estimation. Finally, it is assumed there are more observations than independent variables and there are no exact linear relationships between independent variables.

If these assumptions are believed to hold, estimating parameters by OLS requires minimizing the sum of squared residuals. This amounts to writing equation (13) in terms of estimates:

$$y_t = B_0^{OLS} + B_1^{OLS} x_t + B_2^{OLS} z_t + e_t \quad t = 1, 2, ..., n \tag{16}$$

where the B^{OLS} indicate estimates, and e_t is the error of the estimate each time period, sometimes called the residual. Rearrange this equation and square to get the sum of squared residuals:

$$SSE = \sum_{t=1}^{n} (\hat{y}_t - B_0^{OLS} + B_1^{OLS} x_t + B_2^{OLS} z_t)^2 \tag{17}$$

The equation is squared to remove the influence of negative values. The next step is to choose B_0^{OLS}, B_1^{OLS}, and B_2^{OLS} to minimize this term. This leads to three equations (called first-order conditions) for each time period, which can be solved simultaneously to give the desired estimates. As an

example, the coefficient B_1^{OLS} has the following formula after solving the system of equations:

$$B_1^{OLS} = \frac{\sum_{t=1}^{n}(x_t - \bar{x})y_t}{\sum_{t=1}^{n}(x_t - \bar{x})} \tag{18}$$

The other two coefficients will have a similar formula. This can be generalized to matrices using the same procedure. The formula resulting from minimizing the sum of squared errors in this case is:

$$\hat{B} = (\mathbf{J}'\mathbf{J})^{-1}\mathbf{J}'\hat{y} \tag{19}$$

In this case each of the three coefficients is summarized in the vector $\hat{B}$.

Maximum Likelihood

Another method of estimation used in VAR analysis and more generally is Maximum Likelihood Estimation (MLE). The basic idea behind this approach is to find the coefficient values B^{MLE} which give the greatest probability of obtaining the observed data. To find these coefficient values, one needs to maximize the likelihood function of the sample. Technically, this is the joint probability distribution of the sample, or the probability that the sample values occur together. It is called the likelihood because the likelihood function is interpreted as conditional on the parameters in a statistical model (the linear model above for example), not the sample data.

This bears repeating because it is important in the Bayesian discussion below. Consider flipping a coin 100 times and counting the number of heads. Interpret the number of heads as a random variable (call it X), and any particular outcome of this experiment (realization of the random variable) call x. The random variable will have a distribution based on outcomes, which is the standard probability distribution. It makes sense in this case to ask the probability of getting heads 30 times.

But this is all conditional on the coin being fair. Assume we have a parameter which specifies 1 if the coin is fair and 0 if unfair. The probability distribution is then interpreted as a function of the outcome given a fixed parameter value (in this case 1). The likelihood turns this around. It interprets the same probability distribution as a function of the parameters given fixed outcomes. Here it makes sense to ask the likelihood that the coin is fair (the parameter value), given the observed outcomes. This is a subtle difference but crucial to using and exploiting maximum likelihood estimation, as well as understanding Bayesian techniques.

In order to use MLE in practice for parameter estimation, one needs to derive the likelihood function of the statistical model. To do this, consider the joint density function, or joint distribution of some observations that we assume depend on the parameters in our statistical model, $f(y_1, y_2, ..., y_T|\hat{\theta})$. This distribution is the probability of observing the values of y (the entire sequence) given the parameters of our model ($\hat{\theta}$), the joint probability.

Suppose we are interested in estimating a time series process using MLE, where the errors are Gaussian white noise (i.e. normally and identically distributed and uncorrelated):

$$y_t = Ax_t + \epsilon \tag{20}$$

Each of the outcomes y_t are independent over time, as are the errors, so the joint probability density can be simplified using this independence: $f(y_t, y_{t+1}, ..., y_{t+k}|\hat{\theta}) = f(y_1|\hat{\theta})f(y_2|\hat{\theta})f(y_3|\hat{\theta})...f(y_T|\hat{\theta})$. Because the y_t will inherit the normal distribution of the errors, the joint probability distribution is normal and its formula can be used. The joint distribution becomes:

$$f(y_1, y_2, ..., y_T|\hat{\theta}) = \prod_{t=a}^{T} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) exp\left(\frac{\epsilon_t^2}{2\sigma^2} \right) \tag{21}$$

The right-hand side of this distribution is the formula for a normal probability distribution with the error terms inserted, and σ^2 is the variance of the errors. It is a product of the distributions each for time period because the errors are independent. The next step is to realize that the likelihood function, the likelihood of the parameters given the outcomes, is the same as this joint distribution. That is: $L(\hat{\theta}|y_1, y_2, ..., y_T) = f(y_t, y_{t+1}, ..., y_{t+k}|\hat{\theta}) = f(y_1|\hat{\theta})f(y_2|\hat{\theta})f(y_3|\hat{\theta})...f(y_T|\hat{\theta})$. The final step is to rewrite this by substituting in for the error terms:

$$L(\hat{\theta}|y_1, y_2, ..., y_T|) = \prod_{t=a}^{T} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) exp\left(\frac{-(y_t - Ax_t)^2}{2\sigma^2} \right) \tag{22}$$

One can then choose A and σ^2 to maximize this expression. This tells us the parameter values which make the sample most likely.

Application of MLE in the time series context when there are lags of the endogenous variable on the right-hand side, as in $ARMA$ processes, is a bit more complicated than described above. To illustrate, suppose we are interested in estimating an $AR(1)$ process using MLE, where the errors are Gaussian white noise (i.e. normally and identically distributed and uncorrelated):

$$y_t = \rho y_{t-1} + \epsilon_t \tag{23}$$

The joint probability of these outcomes can be written using Bayes theorem as $f(y_1, y_2, ..., y_T|\hat{\theta}) = f(y_1|\hat{\theta})f(y_2|y_1, \hat{\theta})f(y_3|y_2, y_1, \hat{\theta})...f(y_T|y_{T-1}, ..., y_1, \hat{\theta})$. One must assume that the first observation does not depend on others, then each successive observation depends on the one before, which depends one the one before, and so on. Notice how this differs from the case outlined above where the outcomes were independent over time. Because the y_t will inherit the normal

distribution of the errors, the joint probability distribution is normal and its formula can be used.

In this illustration above, the distribution begins at $t = 1$ and then proceeds to $T = t$. But it is often the case that data available at some initial period are not available. One option is to start at some period p and move until the period T. The joint desnity is then:

$$f(y_1, y_2, ..., y_T|\hat{\theta}) = f(y_1, y_2, ..., y_p|\hat{\theta})f(y_{p+1}|y_p, \hat{\theta})f(y_{p+2}|y_{p+1}, y_p, \hat{\theta})...f(y_T|y_{T-1}, ..., y_p, \hat{\theta}) \quad (24)$$

or

$$L(\hat{\theta}|y_t, y_{t+1}, ..., y_{t+p}) = f(y_1, y_2, ..., y_p|\hat{\theta})f(y_{p+1}|y_p, \hat{\theta})f(y_{p+2}|y_{p+1}, y_p, \hat{\theta})...f(y_T|y_{T-1}, ..., y_p, \hat{\theta}) \quad (25)$$

This joint distribution is also the likelihood function and is called the exact likelihood. The first term on the right-hand side is called the marginal likelihood for the initial values. The idea is that the first p outcomes are taken as given. The remainder of terms form the conditional likelihood. Often, only the conditional likelihood function is used with respect to MLE in the time series context. This is because the marginal likelihood is usually a non-linear function of the parameters, and its use greatly complicates the calculations.

In evaluating the conditional distributions, we know that they will have the same distribution as the errors, which is normal in this case. The formula for the normal distribution can be applied to each conditional distribution for each error term. One can then substitute out the error terms to replace them with the $AR(1)$ process and proceed in the usual way be maximizing the likelihood function. The main point is that autocorrelation does alter the MLE procedure, but using conditional MLE can simplify this complication. This same procedure can be used with VARs as well.

One issue that arises in more complicated situations is that some values of the series are unknown. This may be true of means or variances of the error, or certain observations of the endogenous variables may not be available. This problem can be overcome by putting the general time series into state space form. Once this is done a tool known as the Kalman filter can be used to find these values. This is briefly discussed in the section on State Space Models.

The major drawback to MLE estimation is that specifying a joint probability distribution requires assuming a certain distribution for the error terms. Most econometricians assume that the errors are normally distributed. If this normality assumption is made for the error terms, then the OLS estimator (B^{OLS}) is equivalent to the ML estimator (B^{MLE}). Thus with normal errors, estimating a VAR using OLS is the same as estimating it using maximum likelihood, or using conditional MLE in a time series

context with autocorrelation. While the normality case is the simplest, calculation of the MLE estimator can be much more difficult if the first-order conditions from the maximization problem are non-linear.

Criteria for Estimators

OLS is the most common way to estimate coefficients in a VAR, or in regressions more generally. This is because, when the OLS assumptions are met, it is considered to be the Best Linear Unbiased Estimator (BLUE). In the cases where the assumptions are met, the OLS estimator gives the best tradeoff between unbiasedness and efficiency, or the lowest Mean Square Error (MSE). Both unbiasedness and efficiency of estimators are based on properties of their sampling distributions.

The sampling distribution of an estimator gives the frequency with which values of that estimator occur contingent on the error terms. To illustrate, consider the following representation:

$$y_t = B_0 + B_1 x_t + B_2 z_t + u_t \quad t = 1, 2, ..., n$$

Suppose we postulate a way to estimate the B coefficients in this representation (OLS is only one of many possible ways to do this). To make this calculation, the values of the independent variables and error terms for the sample time period will be needed. Notice that while the values of the independent variables are fixed, the values of the error term are random and will likely change with each draw. This is the property that is exploited to generate a sampling distribution.

The sample data and one draw of the error terms can be used to generate an estimate of the B coefficients (these are generally labeled B^{OLS} to denote that they are estimates). The same sample data in combination with another draw of the error terms can generate another estimate of B^{OLS}. This process can be repeated continually, giving values of the estimator for the B^{OLS} associated with each draw. Once this has been done many times, a sampling distribution for the estimator can be generated. This is a histogram which plots the value of the estimator on the horizontal with the frequency of occurrence on the vertical. See Kennedy (2008) for more on sampling distributions.

If there are two estimators, would you prefer to produce your estimate of B by reaching blindly into the sampling distribution of B^{OLS} or the other estimator B^{OLS*}? The answer usually depends on how close on average each estimator is to the "true" value of the estimate (the bias), as well as how far on average these draws fall from the true value (the variance). The MSE is a summary statistic which weights the bias of each estimator's sampling distribution against its variance (the efficiency). If the assumptions of OLS are met it is preferable to all other linear estimators on these grounds. See Kennedy (2008) for more on this point.

An important point to note is the interpretation of the BLUE criteria. The fact that OLS is unbiased and efficient in a particular context means that if repeated sampling could be undertaken an infinite

number of times (that is, redrawing the error terms), the sampling distribution of the OLS estimator has these properties. Unfortunately an infinite number of draws are not available so this is not a useful interpretation. It may be better to ask: given that I only get one draw, what properties would I want the estimates to have? It seems that unbiasedness and efficiency would be desirable in this case as well.

State Space Representation

The univariate models described above (and the VARs outlined below) can all be put into a common form, called the state space representation. This is convenient representation to work with when in the time series context for several reasons. Importantly, it provides a common structure for most time series models. In fact, all ARMA models have a state space representation. Using this representation can also help in estimating models where outcomes are not independent over time, those where the coefficients vary over time, when the error variance varies over time, or combinations of these.

The general idea behind these models is that an observed time series (y_1,.....,y_T) depends upon a possibly unobserved state vector ($\hat{z}_t$) and this state is driven by a stochastic process. The relation between the observed variable and the (possibly) unobserved state is given by the measurement equation:

$$y_t = \mathbf{H}_t \hat{z}_t + \hat{v}_t \tag{26}$$

where $\mathbf{H}_t$ is a coefficient matrix which can change over time and $\hat{v}_t$ is a vector of observation errors, typically assumed to be white noise processes. The state is assumed to vary according to a transition equation:

$$\hat{z}_t = \mathbf{B}_{t-1} \hat{z}_{t-1} + \hat{w}_{t-1} \tag{27}$$

where $\mathbf{B}_t$ is a coefficient matrix which can change over time and $\hat{w}_t$ is a vector of white noise error processes. As an example, the $ARMA(2,1)$ process:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + e_t + e_{t-1} \tag{28}$$

can be written in state space form with measurement (or observation) equation:

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} y_t \\ A_2 y_{t-1} + e_t \end{bmatrix} \tag{29}$$

where $\hat{z}_t = \begin{bmatrix} y_t \\ A_2 y_{t-1} + e_t \end{bmatrix}$ and $\mathbf{H}_t = \begin{bmatrix} 1 & 0 \end{bmatrix}$. The transition (state) equation can be written as:

$$\begin{bmatrix} y_t \\ A_2 y_{t-1} + e_t \end{bmatrix} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ A_2 y_{t-2} + e_{t-1} \end{bmatrix} + \hat{e}_t \tag{30}$$

where $\mathbf{B}_{t-1} = \begin{bmatrix} A_1 & 1 \\ A_2 & 0 \end{bmatrix}$.

As mentioned above, any $ARMA$ or $VARMA$ model can be put into this form. State space models are usually evaluated using MLE, so the distributional assumptions of the error terms are important. To illustrate why the state space representation is useful, consider a general time series process:

$$y_t = \rho y_{t-1} + \gamma x_t \epsilon \tag{31}$$

This is a combination of the two examples from the MLE section above. To find the MLE estimates of this process, notice that y_t is correlated over time and it also is a function of an exogenous variable x_t. The joint probability of these outcomes can be written as
$f(y_1, y_{t+1}, ..., y_T | x_t, \hat{\theta}) = f(y_1|x_1, \hat{\theta}) f(y_2|y_1, x_2, \hat{\theta}) f(y_3|y_2, y_1, x_3, \hat{\theta}) ... f(y_T|y_{T-1}, ..., y_1, x_T, \hat{\theta})$. The difference from above in the $AR(1)$ case is that outcomes are also conditional on x_t. The likelihood function in general form becomes:

$$L(\hat{\theta}|y_1, y_2, ..., y_T, x_1, x_2, ..., x_T) = f(y_1|x_1\hat{\theta}) \prod_{t=2}^{T} f(y_t|y_1, ..., y_{t-1}, x_t, \hat{\theta}) \tag{32}$$

This can be estimated by MLE, but conditional MLE is easier. What if some of the observations of x are missing? In this case putting the process into a state space representation will allow use of the Kalman filter, which can forecast the missing values of x. This set-up can also be used if one wants to assume that parameters are time-varying. For example, if one believes that the ρ in the $AR(1)$ above is not stable over time. In this case the coefficients are considered to be the state and the outcomes of the time series process are taken as given. As before, an assumption is made about how the state is distributed over time, and the Kalman filter is applied to forecast the state values. Similar logic applies if the variance of the errors is believed to vary over time. Additional details on the Kalman filter and state space representations can be found in Lutkepohl (2007).

Bayesian Techniques and Terminology

Bayesian techniques are also used in VAR analysis along with the classical techniques discussed to this point. Bayesian techniques are particularly important when using VARs for forecasting, or conducting

policy analysis with VARs identified with sign restrictions, or with time-varying volatility. Because each of these latter techniques are covered below, this section provides a brief summary of basic Bayesian terminology and techniques.

Overview

The Bayesian approach is fundamentally different than the classical approach to econometrics. Unlike its conventional counterpart, Bayesian analysis posits a subjective notion of probability, in that the beliefs of modeler are incorporated into model estimation. The differences can best be illustrated through an example which follows Kennedy (2008). Suppose we are interested in estimating the value of an unknown parameter B.

Under the classical approach the data and errors produce a point estimate (B^{OLS}) of B. This estimate has an associated sampling distribution as was described above, which gives the frequency of estimates of B^{OLS} that would be produced in hypothetical repeated samples. This is a distribution of the estimates, not B itself. Any particular estimate using the classical approach is viewed as a random drawing from this sampling distribution. The use of this point estimate is justified by appealing to the properties of its sampling distribution (i.e. unbiasedness, efficiency, etc.). The related confidence interval around this point estimate also has a repeated sampling interpretation, not a probabilistic one. For example, a 95% confidence interval around the point estimate means that if the interval was estimated over and over again, 95% of the time it would cover the true value of B.

The Bayesian approach begins from the proposition that there is only one sample, and hypothetical repeated samples are not relevant. Instead of the point estimate and associated confidence interval, Bayesian estimation produces what is called a posterior distribution or posterior density function. This distribution can be interpreted as the probability that the true value (not the estimate, so this is not a sampling distribution) of B lies between certain points. Think about this. There is no objective point estimate for the true value, rather it is subjective, and summarized in the posterior distribution. In other words, the parameter B is a random variable. This is a very different view of probability and requires some further explanation.

The Bayesian incorporates the subjectivity of the researcher into output of the posterior distribution through the use of a prior distribution. This distribution is the odds a researcher would place on the value of the parameter before viewing the data. This is subjective and can certainty differ between different people. This prior distribution is combined with the likelihood function to yield the posterior distribution using Bayes Theorem. A good way to interpret the posterior distribution is in terms of hypothetical bets. The probabilities give the odds that a researcher would place on the value of the true parameter. For example, if one is willing to bet at even odds that the value of a parameter lies between 2.6 and 2.7, then that person thinks the probability that B lies in this range is greater than 50%. If the opposite bet is taken, the person thinks it is less than 50%.

It bears repeating that the output from a classical estimation procedure is a point estimate with an associated confidence interval. These tell us that if there was repeated sampling, then a certain percent of the time the confidence interval of the estimate would contain the true value of the parameter. The posterior distribution says something very different. It gives the odds that a researcher should take when placing bets on the true value of B. Another way of thinking about the posterior distribution is that it represents a prior distribution modified by the data.

The basic Bayesian framework is conceptually simple. Take two random variables, A and B. The rules of probability imply:

$$p(A, B) = p(A|B)p(B)$$

where $p(A, B)$ is the joint probability of A and B occurring, $p(A|B)$ is the probability of A occurring conditional on B having occurred (the conditional probability of A given B) and $p(B)$ is the marginal probability of B (or just the probability of B by itself). This can also be written as:

$$p(A, B) = p(B|A)p(A)$$

Now these two equation can be set equal to one another and rearranged to give Bayes' theorem:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

This shows the conditional probability of B given A. Interpret A as the sample data and B as the parameters in a statistical model. This formula then gives the posterior distribution: the probability of the parameters (B) given the data (A). How to derive the posterior distribution? It is common practice to drop the denominator because this does not include the parameters. This means that the posterior is proportional to the prior $p(A)$ times the likelihood $p(A|B)$, or:

$$p(B|A) \propto p(A|B)p(B)$$

OLS using Bayesian Techniques: The Univariate Case

Using Bayesian techniques in combination with OLS helps to illustrate some important differences. For generality, Bayesians work with the likelihood function of the linear regression model, as this allows for easing of standard assumptions. For exposition, assume a simple regression model without an intercept as in Koop (2003):

$$y_t = Bx_t + e_t \tag{33}$$

where y is the dependent variable each time period, x_t the independent variable, and B the OLS estimate. The errors, e_t are assumed to be distributed normally, with mean zero, variance σ^2, and e_i and e_j are independent of each other for $i \neq j$. In other words, the errors are i.i.d N(0,σ^2). Also assume that the x_t are independent of the e, with a probability density function $p(x_t|\lambda)$, where λ is a vector of parameters that does not include B. This last assumption rules out correlation between the errors and independent variables. It also means that x is an exogenous variable.

At this point the Bayesian proceeds as one might in MLE, but also using Bayes rule. The first step is to recall that the element of interest for the Bayesian is a posterior distribution of B. That is, a probability distribution of the outcomes of B based on a prior distribution and the likelihood implied by the linear regression model. First look again at Bayes Rule, modified for the current example:

$$p(B, \sigma^2|y_t) \propto p(y_t, x_t|B, \sigma^2, \lambda)p(B, \sigma^2)$$

Notice that the posterior distribution is the joint probability distribution of B and σ^2. The unknown parameters include the variance of the errors. Similarly, the right-hand side of Bayes rule has two joint distributions. The first is the likelihood function, which is the joint probability function of y and x conditional on the parameters. The second is the prior distribution, which is the joint distribution of B and σ^2 prior to observing the data. These joint distributions are not the ultimate objects of interest. We would like the individual posterior distributions of B and σ^2, and it would also be ideal to insert the likelihood and prior distributions in terms of only one variable. This is easiest to see if beginning on the right-hand side.

As is stands, the likelihood function is a joint probability distribution. However, because it is usually assumed that the x_t are exogenous (see the second assumption listed above), y_t and x_t are independent. This means that the likelihood can be rewritten using the laws of probability:

$$p(y_t, x_t|B, \sigma^2, \lambda) = p(y|x_t, B, \sigma^2)p(x_t|\lambda)$$

Because x is exogenous, it depends only on parameters not incorporated into the current model. This also allows separation of the joint distribution, so that y depends only on the exogenous variables and the model parameters. The standard procedure is to drop the distribution of x because it is not usually of interest. At this point standard MLE techniques can be used to derive the likelihood function [$p(y|x, B, \sigma^2)$].[7] The likelihood function is written as a function of B and σ^2, the unknown parameters.

[7] Namely, the likelihood function is normal because the errors are assumed to be normal. The definition of the normal density then forms the basis for the likelihood function, which can be modified from this point.

A similar procedure is used to separate out the joint prior distribution. The coefficient of the regression and the variance of the error distribution are independent (by assumption 1 above), so using the laws of probability the prior can be written:

$$p(B, \sigma^2) = p(B|\sigma^2)p(\sigma^2)$$

The Bayesian must specify a prior for both the B value and the variance of the error terms. At this point experience and subjectivity become very important in Bayesian estimation. There are many different distributions that one might specify for either of these coefficients, which to choose? One method is to choose those that make computation easier, which is the track we follow here. Conjugate prior distributions are ones which, when combined with the likelihood function, yield posterior distributions of the same family. The most common example of such a family is the Gaussian one (i.e. the normal distribution and its many variants). A natural conjugate prior has the additional property that it has the same functional form as the likelihood function.

A common choice to obtain a natural conjugate prior, is to specify the distribution of B to be normal and that of σ^2 to be gamma. The product is a normal-gamma distribution, which requires specification of the respective means and variances of the normal distribution and the gamma distribution respectively. The right-hand side of Bayes rule is now fully characterized:

$$p(B, \sigma^2|y_t) \propto p(y_t|x, B, \sigma^2)p(x|\lambda)p(B|\sigma^2)p(\sigma^2)$$

Given the assumptions on distributions made on the right-hand side, the joint posterior distribution will be normal-gamma. Even with the simplifications, this joint distribution is difficult to calculate. However, this can be simplified further by narrowing focus. Often, the object of interest is the posterior mean of B along with its variance [technically $E(B|y)$ and $var(B|y)$]. This can be calculated as:

$$E(B|y_t) = \int \int Bp(B, \sigma^2|y_t)d\sigma^2 dB$$

There are various methods available for this calculation. Once the mean of the posterior is known, the variance can be backed out as well (given a Gaussian distribution). To summarize, the output of OLS estimation by Bayesian methods is a posterior mean and variance. However, the interpretation is completely different. The mean and variance output from the Bayesian estimation are random variables, those from classical OLS are interpreted as occurring in repeated samples.

Tradeoffs

The Bayesian approach has some clear advantages over its classical counterpart. Importantly, Bayesians are explicit about beliefs via the prior distribution. Classical techniques do not allow this discretion, and it may show up in ad-hoc fashion. Bayesian analysis does not depend on repeated sampling or asymptotic theory. It is more closely aligned with how people think and interpret results. Bayesian analysis can also incorporate additional information quite easily. The main disadvantage to the Bayesian approach is its computational complexity. The theory may be simple but it is not easy to compute the relevant distributions. Some also object to its subjectivity. How can a parameter value be a random variable? Finally, choosing a prior is not an easy task.

Forecasting

Forecasting involves a set of hypothetical statements about future aggregate developments, such as the evolution of output or prices. A popular use of VARs is in generating forecasts, and a brief summary of the basic terminology and issues is reviewed here. This section follows Carnot et al. (2011) and Enders (2010).

Definitions

Conditional forecasts are based on specific assumptions regarding the behavior of agents. At EIA, conditional forecasts are often termed short-term projections. These are forecasts of less than two years which assume that current policy does not change, i.e. they are conditional on current policy. Alternative forecasts are those which describe the most likely scenario. These are often called unconditional forecasts in the literature. At EIA, these are termed short-term forecasts, because they assume that current policy can change over the next two years. Long-run EIA projections are usually alternative forecasts, and these look out more than two years. In what follows the term forecast is used generically as a statement about future developments, without reference to either the policy regime or time frame.

Methods

There are four primary methods of forecasting. Subjective methods are based exclusively on the intuition, experience, and judgement of the forecaster. Indicator-based methods use information in various advance indicators to anticipate movements in certain variables, particularly to detect turning points. Time series models can be either univariate or multivariate and are based on the statistical properties of the series under consideration. There are many different techniques, but $ARMA$ and VAR models are popular examples. Structural models feature causal relationships between different variables which are based on economic theory. The model delivers a forecast of the endogenous variables, and is based on more than the statistical properties of each series.

Properties of Time Series Forecasts

The properties of time series forecasts are easiest to illustrate in an $AR(1)$ framework. The following example follows Enders (2010). Consider an $AR(1)$:

$$y_t = a_0 + a_1 y_{t-1} + \epsilon_t \tag{34}$$

Updating this one period gives:

$$y_{t+1} = a_0 + a_1 y_t + \epsilon_{t+1} \tag{35}$$

If the coefficient values are known, one can forecast y_{t+1} conditional on the information available at t:

$$E_t y_{t+1} = a_0 + a_1 y_t \tag{36}$$

In this cast E_t is short-hand for the conditional expectation of y_{t+1} given the information at t, or $E_t y_{t+j} = E(y_{t+j}|y_t, y_{t-1}, ..., \epsilon_t, \epsilon_{t-1}, ...)$. By substituting in for y on the right-hand side of this equation a conditional expectation can be generated j periods ahead. If this is continually done, the general formula for a j-step ahead forecast is given by:

$$E_t y_{t+j} = a_0(1 + a_1 + a_1^2 + ... + a_1^{j-1}) + a_1^j y_t \tag{37}$$

This forecast function expresses the forecasts based on the information available at t. Taking the limit of this j-step ahead forecast as j goes to infinity, and assuming the process is stationary leads to the conclusion that $E_t y_{t+j} \to \frac{a_0}{1-a_1}$ as $j \to \infty$. It is also the case that the conditional expectation of any stationary $ARMA$ process will converge to the unconditional mean as j goes to infinity.

The forecast errors are defined as the difference between the observed value and the forecasted value. Specifically, the j-step ahead forecast error when forecasting from time t $[e_t(j)]$ is defined:

$$e_t(j) \equiv y_{t+j} - E_t y_{t+j} \tag{38}$$

Beginning at some arbitrary point t the j-step ahead forecast error can also be written as:

$$e_t(j) = \epsilon_{t+j} + a_1 \epsilon_{t+j-1} + a_1^2 \epsilon_{t+j-2} + ... + a_1^{j-1} \epsilon_{t+1} \tag{39}$$

The mean of this error is zero (because the mean of each ϵ is zero), so the forecasts are unbiased estimates of each y_{t+j} (because the expected forecast error is zero). Although there is no bias, there is

variability in the forecasts which can be summarized by the variance:

$$var[e_t(j)] = \sigma^2[1 + a_1^2 + a_1^4 + ... + a_1^{2(j-1)}] \tag{40}$$

As would be expected, the variance of the error rises with the time horizon of the forecast. These are general results for the univariate case with only one lag, but each can be expanded to incorporate more lags, as well as the VAR case, which is discussed in the main text.

Evaluation

The evaluation of forecasts can be either in-sample or out-of-sample. In-sample evaluation consists of using the first t observations of a sample to compute the respective coefficients. The modeler can then calculate the error for the remainder of observations in the sample, say $x + 50$. One method is to use the coefficients from the x observations to calculate all of the errors. The other is to use the coefficients for the x observations to calculate the forecast error with respect to $x + 1$, then re-estimate the coefficients with the $x + 1$ observations, and calculate the forecast error with respect to observation $x + 2$. This procedure can be done for the remainder of observations in the sample. Out-of-sample forecast errors are calculated with observations which are not currently in the sample, and usually become available only over time.

Whichever method is used, the forecaster must pick a way to summarize the forecast errors. There are multiple methods available for this task, and the choice of the best one will depend on the purpose of the forecast. For example, the accuracy criterion used for forecasting turning points in the business cycle will generally be different than one used for forecasting GDP growth at various horizons.

A popular choice to summarize forecast accuracy is mean absolute deviation (MAD). The is the average of the absolute values of the forecast errors. This is sometimes also called mean forecast error (MAE), and is appropriate when the cost of forecast errors is proportional to the absolute size of the forecast error. It is sensitive to scaling. Root mean square error (RMSE) is also sensitive to scaling, and is the square root of the average of the squared values of the forecast errors. This measure weights large forecast errors more heavily than small ones.

Mean absolute percentage error (MAPE) is not sensitive to scaling, and is the average of the absolute values of the percentage errors. It is appropriate when the cost of the forecast error is more closely related to the percentage error than to the numerical size of the error. Another measure is the correlations of forecasts with actual values. The percentage of turning points criteria is a $0/1$ measure which summarizes if turning points were forecast correctly. Each of these has its variations, and there are other methods as well which may be used for specialty forecasts.

A General VAR and Associated Representations

Vector autoregressions are a natural extension to univariate autoregressions in the context of time series. Instead of assuming that only a variable's own lags impact its current value, VARs allow for the lags of multiple variables to impact each other. This is very useful in empirical macroeconomics because many variables are interrelated, and VARs provides a means to asses those interrelations.

Below a basic VAR representation is outlined. For illustration, a two-variable VAR with one lag is fully characterized. This is then expanded to the general case. A discussion of the requirements and issues with respect to estimation of a VAR are outlined, and the vector moving average (VMA) representation is then highlighted in the context of impulse responses, variance decompositions, and historical decompositions. The use of these methods is often referred to as innovation accounting, and they are important for conducting policy analysis with VARs. This section follows Enders (2010) and Lutkepohl (2007), while also drawing examples from Kilian (2009) and Cochrane (1994).

A Basic VAR

In a vector autoregression, the variables of interest (endogenous variables) form a vector. It is assumed that each of these endogenous variables impacts the others, possibly simultaneously. This relationship is summarized in the structural representation of a VAR, which postulates that this vector of endogenous variables can be approximated by a vector autoregression of order p. For the two variable, first-order case this reads:

$$b_{11}y_t = b_{10} - b_{12}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \epsilon_{1t} \tag{41}$$

$$b_{22}z_t = b_{20} - b_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \epsilon_{1t} \tag{42}$$

It is important to fully understand each element of these two equations. The first point to note is that the existence of such a linear relationship is in itself an assumption. One should always ask if this is a reasonable assumption. In this case, the endogenous variables are y and z, say GDP and the oil price. Notice how the current values of each endogenous variable are allowed to impact the other, making this a structural representation. Because there is only one lag of each variable in either equation, the system is first-order. The coefficients summarize the impact of each variable on the other. For example, $-b_{21}$ is the contemporaneous impact of a unit change in y_t on z_t, and $-b_{12}$ is the contemporaneous impact of a unit change in z_t on y_t.

The ϵ_t may be referred to as structural innovations or structural shocks or structural disturbances, and are also termed residuals as well. When using a VAR to study the impact of one variable on another (sometimes broadly referred to in the literature as policy analysis) these are the primary objects of interest. They are assumed to be a white noise processes. Due to these assumptions, the shocks

represent unexpected movements in either y_t (ϵ_{1t}) or z_t (ϵ_{2t}). Technically, they do not have to be specifically related to a specific variable, although they are often interpreted in this way. Rather, they are the causes of unexpected movements in the value of that variable which are unpredictable and uncorrelated with other endogenous variables or innovations.

To repeat, the structural innovations are unpredictable and uncorrelated with either previous realizations at any lead/lag or realizations of other structural innovations at any lead/lag. Therefore, a shock to one structural innovation which changes the value of an endogenous variable is the item of interest in analysis which seeks to isolate the impact of various policies. Consider a one unit increase in ϵ_{1t} in the equations above. This unexpected and unpredictable movement will increase y_t, by equation (41). Equation (42) shows that this will result in an increase in z_t as well. Because the structural innovation ϵ_{1t} could not be predicted, neither y_t nor z_t could have moved in expectation of this change. And because ϵ_{1t} is uncorrelated with anything else, there is no other factor causing the shock to occur. Taken together with the fact that it works through y_t, this implies that the movement in z_t following the one unit increase in ϵ_{1t} can be interpreted as the impact of unexpected movements in y_t on z_t. Using the responses of endogenous variables to structural innovations provides a way to extract the impact of one variable on another.

A crucial point is that the structural innovations are assumed to be represented by the residuals. Do they really exist? Are there such things as unexpected movements uncorrelated with anything? Ultimately this question is difficult (maybe impossible) to answer and depends on the beliefs of the modeler. But it explains some of the skepticism which VARs have received when used for policy analysis, as this requires a very specific interpretation of the residuals. Another point to notice is that the coefficient values play no role in assessing the impact of shocks on variables in the structural system. Unlike in traditional autoregression analysis, the coefficient estimates are not particularly important when using VARs in this form. However, we will see below that they are important when estimating the reduced form and then using impulse response analysis, as well as in forecasting.

VARs are generally estimated using OLS, which requires transforming the structural equations to reduced form. This is because estimation by OLS requires the right-hand side variables be uncorrelated with the error term. A quick look at equations (41) and (42) shows this is not the case if current values of either endogenous variable remain on the right-hand side. For the two variable, first-order VAR above the reduced form representation is given by:

$$\underbrace{\begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix}}_{\mathbf{B}} \underbrace{\begin{bmatrix} y_t \\ z_t \end{bmatrix}}_{\hat{x}_t} = \underbrace{\begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix}}_{\hat{\Gamma}_0} + \underbrace{\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}}_{\mathbf{\Gamma}_1} \underbrace{\begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix}}_{\hat{x}_{t-1}} + \underbrace{\begin{bmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{bmatrix}}_{\hat{\epsilon}_t} \tag{43}$$

Notice here that both b_{11} and b_{22} have been normalized to equal one, which is standard. More

succinctly:

$$\mathbf{B}\hat{x}_t = \hat{\Gamma}_0 + \mathbf{\Gamma}_1\hat{x}_{t-1} + \hat{\epsilon}_t \tag{44}$$

Pre-multiplication by $\mathbf{B}^{-1}$ yields the VAR in reduced (or standard) form:

$$\hat{x}_t = \mathbf{A}_0 + \mathbf{A}_1\hat{x}_{t-1} + \hat{e}_t \tag{45}$$

Where $\mathbf{A}_0 = \mathbf{B}^{-1}\hat{\Gamma}_0$, $\mathbf{A}_1 = \mathbf{B}^{-1}\mathbf{\Gamma}_1$, and $\hat{e}_t = \mathbf{B}^{-1}\hat{\epsilon}_t$. Take a close look at this general representation. The main difference from the structural form is that current values of either endogenous variable are no longer on the right-hand side of the equation. Each matrix has also been transformed by pre-multiplication, including the structural innovations. This is the key point: the reduced form no longer directly represents the structural shocks. Rather, the reduced form is based on a transformation of the structural shocks, namely $\hat{e}_t$. This becomes clearer if written out fully:

$$y_t = a_{10} + a_{11}y_{t-1} + a_{12}z_{t-1} + e_{1t} \tag{46}$$

$$z_t = a_{20} + a_{21}y_{t-1} + a_{22}z_{t-1} + e_{2t} \tag{47}$$

The items of interest for policy analysis, the structural shocks, are no longer directly represented here. They have been replaced by error terms (e_{1t} and e_{2t}). Also notice again how this system has each endogenous variable dependent only on lags of itself and lags of the other endogenous variable. Because of this, the two variable first-order system can be estimated by OLS equation-by-equation to yield coefficient estimates and associated error values. The validity of the estimates, however, depends upon the assumptions underlying OLS holding.

The reduced form system can also be generalized to an arbitrary number of variables (n) and lags (p):

$$\mathbf{C}_0\hat{w}_t = \mathbf{C}_1\hat{w}_{t-1} + \mathbf{C}_2\hat{w}_{t-2} + ... + \mathbf{C}_p\hat{w}_{t-p} + \hat{v}_t \tag{48}$$

Here, $\hat{w}$ is a nx1 vector of endogenous variables, the $\mathbf{C}$ are nxn matrices of coefficients, and $\hat{v}$ is an nx1 vector of errors. This system is a direct generalization of equation (44) above, and all of the same points mentioned above apply. It can be written more compactly using the lag operator:

$$\mathbf{C}(L)\hat{w}_t = \hat{v}_t \tag{49}$$

where $\mathbf{C}(L) = \mathbf{C}_0 - \mathbf{C}_1L - \mathbf{C}_2L^2 - ... - \mathbf{C}_pL^p$ is the autoregressive lag order polynomial.

Examples of VARs

VARs are a very common tool in macroeconomic policy analysis. Two examples of papers which use VARs are presented below, and are referred to and used throughout the remainder of this note. The first is an example of a VAR used to asses the factors that influence oil prices, Kilian (2009). The second is an older paper on the impact of monetary policy on the economy, Cochrane (1994).

Example 1: The Price of Oil, Kilian (2009)

Kilian (2009) attempts to understand and quantify the different factors which impact the price of crude oil. He chooses to use a VAR with three variables: the change in world crude oil production ($\Delta prod$), a measure of global economic activity (rea), and the real price of oil (rpo). According to Kilian (2009), the first two variables have a very important effect on the oil price, and the remainder of the influences come from the price itself. This means that the vector of endogenous variables from equation (45) can be written as:

$$\hat{x}_t = \begin{bmatrix} \Delta prod_t \\ rea_t \\ rpo_t \end{bmatrix}$$

Kilian (2009) explains that including these variables is a logical way to differentiate between the factors which influence the price of oil. The $\Delta prod$ represents the impact of oil supply, while rea represents the impact of demand. Both of these variables are conventionally believed to impact the oil price. The major insight in this approach is that rea is interpreted as the demand for all industrial commodities, not just oil. Thus it represents price rises due to economic growth generally. The final factor influencing the price of oil is demand which is specific to the oil market, possibly due to precautionary demand for inventories. Kilian (2009) argues that rpo captures this specific demand.

Example 2: The Impact of Monetary Policy, Cochrane (1994)

Cochrane (1994) brings together the early literature on the impact of monetary shocks on the economy. Although he uses a variety of models, we will focus on the simplest one here. The VAR in this case has three variables: a measure of monetary policy ($M2$), GDP (y), and the price level (p). The idea is to gauge how the latter two variables respond when there are changes in monetary policy.

The next few sections cover how to quantify and use the results from estimated VARs.

Estimation

OLS is commonly used to estimate an arbitrary reduced form VAR. However, before this can be done five issues must be considered and addressed. The first is which variables to include in the system, the second is the choice of lag length for these variables, the third is whether or not the variables need to be stationary, the fourth issue is if the assumptions of OLS are met, and the final choice is if Bayesian methods are appropriate.

The variables to include in a VAR are usually selected according to some relevant economic model. As with any empirical model, the modeler seeks to incorporate the fewest number of variables in hopes of keeping the model as simple as possible. Even with parsimonious models, one issue that often arises in the case of VARs is that coefficient estimates are insignificant. This is called overparameterization, and occurs because there are so many coefficients to be estimated in a VAR: each endogenous variable has an equation which has lags of every other endogenous variable. With many lags the number of coefficients to estimate can grow quite large. Fortunately, this is usually not considered a problem when using VARs for policy analysis. As mentioned above, the goal in policy analysis is to find (and quantify) the interrelationships between variables, and the coefficient estimates are not helpful in this regard. However, this is a major issue when using VARs for forecasting, and some alternatives are described below in the section on forecasting.

Choosing the appropriate lag length can be more difficult, although there are a variety of tests which can aid in the process. The most common way to proceed in order to preserve the symmetry of the system is to choose the same lag length for all variables in the system. It is also generally accepted when working with macroeconomic variables that the lag-length should be at least one year to capture any seasonality in the data, although some modelers do not follow this practice. The tests of lag length amount to finding the lag length of the model which provides the best fit according to some criterion. The two most popular tests are based on the Akaike Information Criterion (AIC) and the Schwartz Bayesian Criterion (SBC). In general, the SBC tends to penalize additional lags more than the AIC, while the AIC is somewhat more general in weighing the benefits of adding lags versus the costs in terms of model fit. These criterion can be particularly helpful when using VARs for forecasting in choosing which of the overparameterized coefficients to drop.

There are also different ways to proceed when variables in a VAR contain a unit root. If the goal is to uncover the interrelationships between the variables, then the general view is to leave the variables untransformed when estimating the VAR. Transforming the data can throw away information on their interrelationships, which is the object of interest when using VARs for policy analysis. This advice does not hold if the goal is to use a VAR for forecasting, as the coefficients are the items of interest from the estimation.

The fourth consideration is to make sure the assumptions of OLS are met. The first assumption requires each variable to be a linear function of the other endogenous variables, which is implicit in the structural and reduced form representations of a VAR. The second and third assumptions on the error term can be handled by assuming that the structural shocks are white noise. The fourth assumption, which rules out correlation between the error term and right-hand side variables, is the reason why the system must be estimated in reduced form. The final assumption is met by ensuring there are enough observations, and that none of the endogenous variables are exact linear combinations of each other.

When this is done, each equation is estimated one at a time, and all of the assumptions of OLS must

hold. One assumption that commonly fails is that the observations of independent variables can be considered fixed in repeated samples. This fails because the VAR is really a system of simultaneous equations where all of the variables of interest are endogenous. In practice this means that there will be correlation between the reduced form errors in each equation.

When estimating a VAR this complication is usually ignored. This is because each equation has the same variables on the right-hand side, meaning there are no gains from using an alternative method (see Enders (2010)). If the VAR is to be used for forecasting, it is often the case that insignificant parameter values are dropped (a near-VAR). In this case the right-hand side variables are no longer the same either, and seemingly unrelated regressions (SURE) are commonly used to estimate near-VARs.

The final choice relates to the use of Bayesian methods. Particularly in forecasting, it may be advantageous to formally incorporate prior information alongside the data when estimating the coefficients. Bayesian methods can also be used in conjunction with VARs when conducting policy analysis. The drawbacks of this approach are the added computational complexity (which seems to be coming down), and also the choice of how to specify prior distributions. The strengths and weaknesses of either the Bayesian or classical approach are addressed in the sections on policy analysis and forecasting below.

Impulse Responses

An impulse response function is a powerful way to summarize the instantaneous and continuing impact of movements in structural innovations on the endogenous variables in a VAR. It graphically summarizes the impact of a one time, one unit increase of a structural innovation (often called a shock) on each endogenous variable in the system. Figure 1 shows typical impulse response functions from a small VAR, taken from Cochrane (1994).

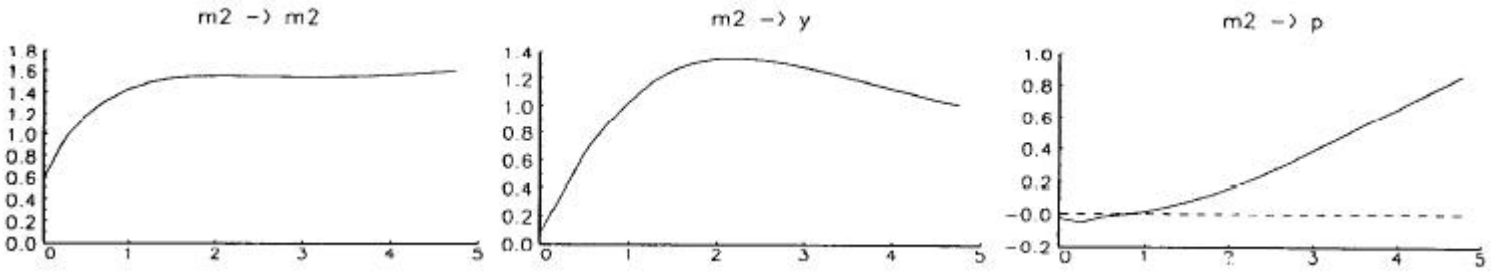


Figure 1: Response of $M2$, GDP (y), and the price level (p) to a one unit innovation to $M2$ from Cochrane (1994). Horizontal axis in years, vertical in %.

The individual plots depict the response of a measure of the money supply ($M2$), GDP (y), and the price level (p) to a one unit innovation in $M2$. That is, they are addressing the impact of monetary shocks (measured by $M2$) on GDP and inflation. The impulse responses of this particular configuration support a standard view of monetary policy: money is non-neutral and impacts output

with lags, and money impacts the price level in the long run.

Deriving the impulse response functions of a VAR requires converting from the vector autoregressive form to a vector moving average (VMA) representation. Recall that any stationary $AR(1)$ process has an $MA(\infty)$ representation. Similarly, a $VAR(1)$ process also has a $VMA(\infty)$ representation. Fortunately, the conversion is not limited only to VARs with one lag. A $VAR(p)$ process can be transformed into a $VAR(1)$ process, which can then yield a $VMA(\infty)$ representation. The basic idea is that the matrices and vectors of the $VAR(p)$ process are grouped together in arrays and matrices, similar to the way in which vectors and scalars are grouped into matrices and vectors when writing OLS in matrix form. See Cochrane (2005) or Lutkepohl (2007) for examples of the procedure.

To illustrate the impulse response functions, begin with the two variable case using equations (46) and (47) from above:

$$y_t = a_{10} + a_{11}y_{t-1} + a_{12}z_{t-1} + e_{1t} \tag{50}$$

$$z_t = a_{20} + a_{21}y_{t-1} + a_{22}z_{t-1} + e_{2t} \tag{51}$$

Putting these only in terms of the errors gives (where $\bar{y}$ and $\bar{z}$ are constants):

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^i \begin{bmatrix} e_{1t-i} \\ e_{2t-i} \end{bmatrix} \tag{52}$$

This equation is derived by continually substituting lagged values of the endogenous variables back into equations (46) and (47). To make this clearer it is helpful to write out a few of the summation terms:

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^0 \begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^1 \begin{bmatrix} e_{1t-1} \\ e_{2t-1} \end{bmatrix} + ... + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^j \begin{bmatrix} e_{1t-j} \\ e_{2t-j} \end{bmatrix} \tag{53}$$

Using this form shows that the matrix of a values summarizes the impact of a change in the respective error term on the endogenous variables. For example, at t a one unit change in e_{1t} has a one unit impact on y_t. Similarly, the impact of a one unit change in e_{1t-1} on y_t is given by a_{11}.

This illustrates the VMA representation of the reduced-form VAR. However, our ultimate interest is in the structural innovations. These responses are captured in the coefficients of the matrix $\mathbf{B}$, which can be derived from the individual error terms using $\hat{e}_t = \mathbf{B}^{-1}\hat{\epsilon}_t$, or:

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} = \frac{1}{1 - b_{12}b_{21}} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{bmatrix} \tag{54}$$

The next step is to substitute this equation in to the moving average representation, equation (52):

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} + \frac{1}{1 - b_{12}b_{21}} \sum_{i=0}^{\infty} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^i \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \begin{bmatrix} \epsilon_{1t-i} \\ \epsilon_{2t-i} \end{bmatrix} \tag{55}$$

This form is very powerful because it directly shows the impact of the structural innovations on each endogenous variable. This become clearer if a matrix is defined to simplify the coefficients:

$$\phi(i) = \frac{\mathbf{A}_1^i}{1 - b_{12}b_{21}} \begin{bmatrix} 1 & -b_{12} \\ -b_{21} & 1 \end{bmatrix} \tag{56}$$

In this notation all of the a coefficients are put into the matrices $\mathbf{A}_1^i$, which is why ϕ is indexed by i, meaning that it is different for each lag. Using this notation, equation (55) can be rewritten:

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11}(i) & \phi_{12}(i) \\ \phi_{21}(i) & \phi_{22}(i) \end{bmatrix} \begin{bmatrix} \epsilon_{1t-i} \\ \epsilon_{2t-i} \end{bmatrix} \tag{57}$$

The coefficients ϕ summarize the impacts of changes in the structural innovations on the endogenous variables. These are often called the impact multipliers. The multiplier of most interest corresponds to the current period t, when $i = 0$. For example, $\phi_{11}(0)$ is the instantaneous impact of ϵ_{1t} on y_t, and $\phi_{12}(0)$ is the instantaneous impact of ϵ_{2t} on y_t. So if we are interested in asking about the instantaneous impact of current period shocks to z_t (due to ϵ_{2t}) on current period y_t, these are summarized by the impact multiplier $\phi_{12}(0)$. Relatedly, these are the instantaneous movements in $M2$ and y in Figure 1 above.

The coefficients $\phi(i)$ are also called the impulse response functions. Plotting these functions gives the impact of the innovations on the endogenous variables. Writing out the summation can make this clearer:

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} + \begin{bmatrix} \phi_{11}(0) & \phi_{12}(0) \\ \phi_{21}(0) & \phi_{22}(0) \end{bmatrix} \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix} + \begin{bmatrix} \phi_{11}(1) & \phi_{12}(1) \\ \phi_{21}(1) & \phi_{22}(1) \end{bmatrix} \begin{bmatrix} \epsilon_{1t-1} \\ \epsilon_{2t-1} \end{bmatrix} + ... + \begin{bmatrix} \phi_{11}(j) & \phi_{12}(j) \\ \phi_{21}(j) & \phi_{22}(j) \end{bmatrix} \begin{bmatrix} \epsilon_{1t-j} \\ \epsilon_{2t-j} \end{bmatrix} \tag{58}$$

Compare this summation with Figure 1. The horizontal scale of Figure 1 begins at period t and moves five periods into the future. Equation 58 gives the value of the variables at t based on error and coefficient values at and before t. The equation is backward looking while the figure begins at t and moves forward. It is standard to plot impulse response functions in this way, and equation (58) can be transformed to show this representation. Suppose both of the structural innovations in this equation increase by one unit at t. The instantaneous responses of y_t to ϵ_{2t} is summarized by $\phi_{12}(0)$. In

equation (58), $\phi_{12}(1)$ is the impact of a one unit change in ϵ_{2t-1} on y_t. This is equivalent to moving forward one period and interpreting $\phi_{12}(1)$ as the impact of a one unit change in ϵ_{2t} on y_{t+1}. Similarly, $\phi_{12}(2)$ is interpreted as the impact of a one unit change in ϵ_{2t} on y_{t+2}, and so on. A plot of the impulse response functions which moves forward such as Figure 1 shows the impact of a current period structural innovation on the value of an endogenous variable into the future.

The discussion to this point has assumed that the structural innovations can be recovered from the underlying structural representation using OLS estimation. This is not an easy or uncontroversial task, and requires additional assumptions. The section on policy analysis covers this issue at length.

A final point with the impulse response functions is that they are constructed using estimated coefficients. Due to this, it is very likely the VAR is overparameterized (see discussion below in sections on policy analysis and forecasting), implying that the impulse responses contain errors. To account for this, and to attempt to quantify the error, it is standard to construct confidence intervals around the impulse responses. These allow for the parameter uncertainty inherent in the estimation, just as t-statistics do in basic regression analysis. Constructing the respective intervals analytically can be complicated, and Monte Carlo procedures are often used.

This is accomplished in multiple steps. The first step is to estimate the respective coefficients of the reduced form VAR using OLS. The error terms from this estimation are saved along with the coefficients. The next step is to draw a random sample of size T (the size of the sample) from a normal distribution to represent the error sequence. At this point we have the estimate coefficients and residuals, along with a random draw of residuals. Now use the random draw of residuals, along with the estimated coefficients, to back out the implied values of the endogenous variables. Then using these implied values, re-estimate the VAR and compute the impulse response functions.

This process is repeated several thousand times. The result will be several thousand impulse response functions, each based on the random draw of residuals leading to implied values for the endogenous variables. The confidence intervals are calculated by dropping the highest and lowest x percent. Visually, a plot would have the estimated impulse response function from the VAR bracketed by the confidence intervals. The highest confidence interval represents, for example, the point below which are 97.5 percent of the impulse responses from the Monte Carlo. The lowest confidence interval represents the point above which are 97.5 of the Monte Carlo impulse responses. The area in-between these two is then the area where 95 percent of the Monte Carlo impulse response fall.

Variance Decompositions

Another tool used to uncover interrelationships between variables in VARs is forecast error variance decomposition (FEVD), or variance decomposition (VD). This tool is used to study properties of forecast errors from a VAR. The forecast error is the difference in the actual value of a process at $t+j$ and its predicted value at $t+j$ made at t. For example, take the VAR in standard form above,

equation (45):

$$\hat{x}_t = \mathbf{A}_0 + \mathbf{A}_1 \hat{x}_{t-1} + \hat{e}_t$$

Suppose the coefficients have been estimated. At t, the expected value of $\hat{x}_{t+1}$ is generated by moving this equation forward one period and taking the expectation, which is written:

$$E_t \hat{x}_{t+1} = \mathbf{A}_0 + \mathbf{A}_1 \hat{x}_t \tag{59}$$

Here E_t is the conditional expectation of $\hat{x}_{t+1}$ at t based on previous values, i.e. $E_t = E_t(\hat{x}_{t+1}|\hat{x}_t, ..., \hat{x}_1)$, and the error term drops out because its expected value is zero by construction. The one-step ahead forecast error is the difference between equation (45) updated one period forward and equation (59), or:

$$\hat{x}_{t+1} - E_t \hat{x}_{t+1} = \hat{e}_{t+1} \tag{60}$$

This can be generalized to give the n-step ahead forecast error:

$$\hat{x}_{t+n} - E_t \hat{x}_{t+n} = \hat{e}_{t+n} + \mathbf{A}_1 \hat{e}_{t+n-1} + \mathbf{A}_1^2 \hat{e}_{t+n-2} + ... + \mathbf{A}_1^{n-1} \hat{e}_{t+1} \tag{61}$$

In this form the forecast error is not very informative. It becomes more useful if written in VMA form based on the structural innovations. The VMA form puts the forecast error completely in terms of the structural innovations. One can quantify the extent to which each structural innovation contributes to the total forecast error. This is one means of assessing how important different structural innovations are for movements in a particular endogenous variable. To put the forecast error into VMA form, first rewrite equation (57) from above in more general notation:

$$\hat{x}_t = \hat{\mu}_t + \sum_{i=0}^{\infty} \boldsymbol{\phi}_i \hat{\epsilon}_{t+n-i} \tag{62}$$

This process can be moved forward n periods, with a forecast error given by $\hat{x}_{t+n} - E_t \hat{x}_{t+n}$. Because the innovations are white noise, the associated n-step ahead forecast error is given by the weighted sum of structural innovations from $t+1$ to $t+n$:

$$\hat{x}_{t+n} - E_t \hat{x}_{t+n} = \sum_{i=0}^{n-1} \boldsymbol{\phi}_i \hat{\epsilon}_{t+n-i} \tag{63}$$

At this point, the goal is compute the contribution of each structural innovation to the forecast error

in each endogenous variable. This information is summarized in the coefficients in ϕ_i. These coefficients are also the impact multipliers used for impulse response analysis. To simplify, consider the n-step ahead forecast error in the two-variable case, using one of the endogenous variables, y:

$$y_{t+n} - E_t y_{t+n} = \phi_{11}(0)\epsilon_{yt+n} + \phi_{11}(1)\epsilon_{yt+n-1} + ... + \phi_{11}(n-1)\epsilon_{yt+1} + \phi_{12}(0)\epsilon_{zt+n} + \phi_{12}(1)\epsilon_{zt+n-1} + ... + \phi_{12}(n-1)\epsilon_{zt+1} \quad (64)$$

This equation explicitly expands equation (63) for one endogenous variable, y. It breaks down the forecast error of y into functions of unexpected movements (innovations) in each variable. It allows us to discuss how much of the forecast error is due to unexpected movements in z versus unexpected movements in y. Why is this important? It allows us to quantify which impacts a variable more. To ease interpretation, this is often put in terms of a variance. Using equation (64), the variance of the n-step ahead forecast error for each endogenous variable can be calculated. The fraction of that variance due to each structural innovation can also be calculated from equation (64). This is because each structural innovation is separated from the others.

Table 1 shows the variance decomposition from the VAR in Figure 1, which comes from Cochrane (1994). The entries are the percent of the n-step ahead forecast error variance due to the column innovations.

	Shock and Horizon											
	1 Qtr.			1 Year			2 Year			3 Year		
Var. of	$M2$	y	p	$M2$	y	p	$M2$	y	p	$M2$	y	p
$M2$	100	0	0	99	1	0	98	0	2	94	1	5
y	1	99	0	32	68	0	70	30	0	82	17	1
p	1	3	96	0	7	92	1	17	83	3	24	73

Figure 2: Variance Decomposition. Table entries are the percent of the forecast error variance of the row variable due to the column innovation at the specified horizons.

In this case n is specified at the top of the table (1 quarter, 1 year, etc.). For example, at one quarter the innovation to $M2$ only accounts for one percent of the forecast error in y. 99% of this forecast error is due to the variance of y itself. By one year, however, this has changed. Now the original innovation to $M2$ accounts for 32% of the forecast error in y. The interpretation is that shocks to money supply impact output with a lag. 32% of the forecast error of output in one year is accounted for by the unexpected movement in $M2$ at t, which means that $M2$ matters for movements in y in this framework.

Historical Decompositions

Historical decompositions are another tool for use with VARs. They decompose the value of an endogenous variable in the VAR at any point in time in terms of only the structural innovations. This allows the analyst to make statements about the relative historical importance of one structural innovation versus another in determining the value of an endogenous variable. Historical decompositions use the fact that the value of an endogenous variable from a VAR at any point in time can be decomposed as the sum of the forecast error and the forecast. To see this begin with equation (63):

$$\hat{x}_{t+n} - E_t\hat{x}_{t+n} = \sum_{i=0}^{n-1} \boldsymbol{\phi}_i \hat{\epsilon}_{t+n-i}$$

Then add the forecast to both sides:

$$\hat{x}_{t+n} = \sum_{i=0}^{n-1} \boldsymbol{\phi}_i \hat{\epsilon}_{t+n-i} + E_t\hat{x}_{t+n} \tag{65}$$

This equation is an identity, as it follows from the definition of the forecast error. The summation on the right-hand side is the portion of $\hat{x}_{t+n}$ due to structural innovations from $t+1$ to $t+n$. Given equation 63, $E_t\hat{x}_{t+n}$ must give the value from the starting point to t because $\hat{x}_{t+n}$ is the sum of all structural innovations from the starting period to $t+n$. This term can be expanded by repeatedly substituting equation (59) into itself in different time periods, which eventually leads to the following general equation:

$$E_t\hat{x}_{t+n} = (\mathbf{I} + \mathbf{A}_1 + \mathbf{A}_1^2 + ... + \mathbf{A}_1^{n-1})\mathbf{A}_0 + \mathbf{A}_1^n\hat{x}_t \tag{66}$$

This gives the expected value at t in terms of coefficient matrices and the current value of $\hat{x}$. The current value of $\hat{x}_t$ in VMA form is given by equation (62):

$$\hat{x}_t = \hat{\mu}_t + \sum_{i=0}^{\infty} \boldsymbol{\phi}_i \hat{\epsilon}_{t+n-i}$$

Putting equation (62) into equation (66) and substituting this into equation (65) yields (where $\hat{\Gamma}$ is a constant of coefficient values):

$$\hat{x}_{t+n} = \sum_{i=0}^{n-1} \boldsymbol{\phi}_i \hat{\epsilon}_{t+n-i} + \sum_{i=n}^{\infty} \boldsymbol{\phi}_i \hat{\epsilon}_{t+n-i} + \hat{\Gamma} \tag{67}$$

The value of the endogenous variables is fully in terms of structural innovations and coefficients. The first summation term on the right-hand side gives the contribution of these innovations from $t+1$ to $t+n$ to the value of the endogenous variables at $t+n$. The second summation on the right-hand side gives their contribution from the starting point to t on the values of the endogenous variable. Although this summation starts in the infinite past, the terms far into the past will have little impact on the value at $t+n$, so it can be truncated.

Figure 2 shows an example of a historical decomposition from Kilian (2009). The figure shows the contributions of three different shocks, or structural innovations on the real price of oil: oil supply, aggregate demand, and oil-market specific demand. The price in constant dollars is on the vertical axis, and time in years on the horizontal.

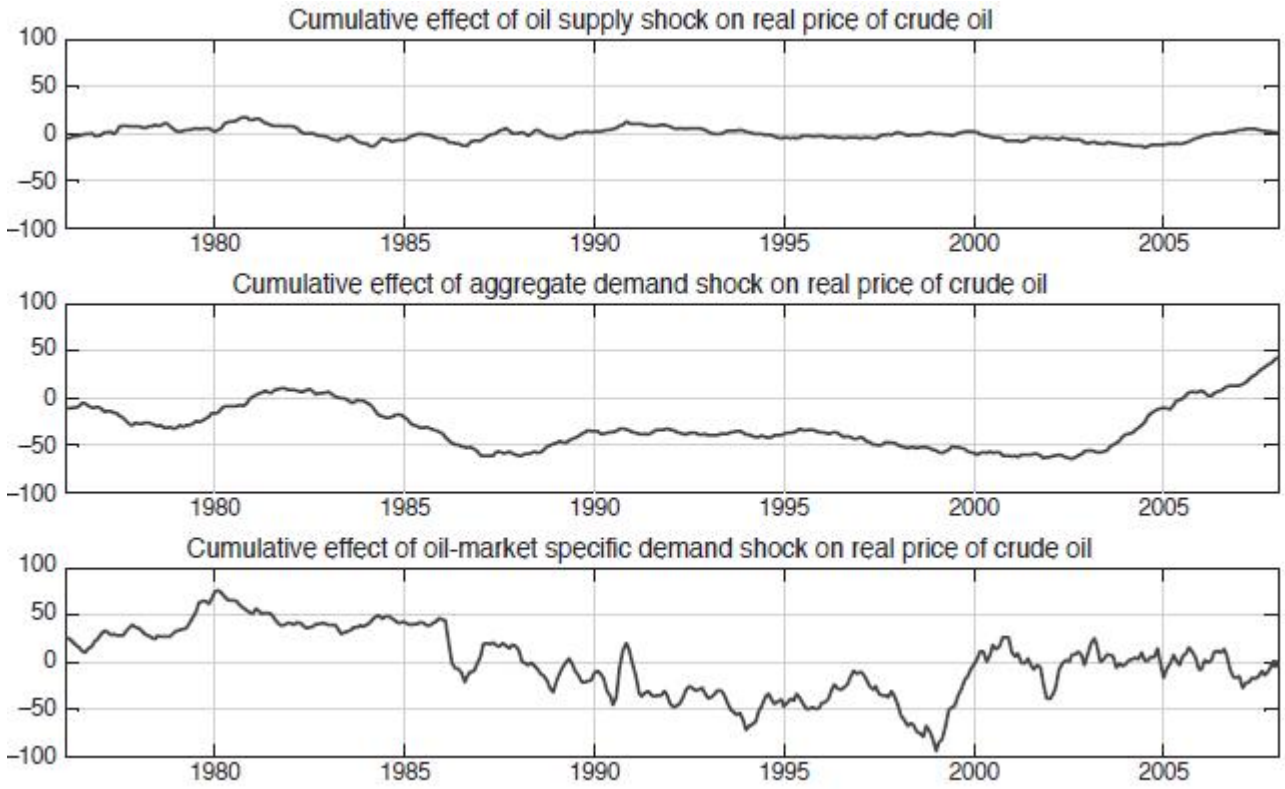


Figure 3: Historical decomposition of the real price of oil in terms of three structural innovations: oil supply, aggregate demand, and oil-market specific demand from Kilian (2009).

The impact of structural innovations can be positive or negative. The total of all three charts for any year must sum to the real price of oil in that year. It is also common to plot each as a fraction of the price. Notice that oil supply shocks do impact the oil price, but seem to have relatively small influence compared with both demand side innovations. And the variance of the oil price seems to come mostly from the oil-market specific demand shocks.

Basic Policy Analysis with VARs

There is an enormous literature on the uses of VARs for policy analysis. The basic goal of this literature is to quantify the impact of changes in one variable on another. In the context of VARs this boils down to isolating the structural innovations of one variable and understanding how its movements impact another endogenous variable. This understanding or quantification is done by using impulse responses, variance decompositions, or historical decompositions, which collectively is called innovation accounting.

The previous section described and detailed VARs in general, this section focuses on policy analysis. It begins with an explanation of why it is difficult to extract the structural innovations of interest from the reduced form representation. The next section describes in general how the reduced form can be used to recover structural innovations, details the most common method called recursive identification, and addresses any resulting changes to the innovation accounting procedures. Policy analysis using Bayesian estimation is briefly discussed in the final part, along with the pros and cons of using VARs for policy analysis. The material in this section comes mainly from Enders (2010) and Lutkepohl (2007), while also drawing examples from Kilian (2009) and Cochrane (1994).

The Identification Issue

When using VARs for policy analysis, the items of interest are the structural innovations. Specifically, we are interested in assessing the interrelationships between variables, often quantified through impulse response analysis, variance decompositions, or historical decompositions. The individual structural shocks provide a basis for summarizing these interrelationships because they represent unexpected movements in endogenous variables that are uncorrelated with the innovations in other endogenous variables.

The previous section explains why only a reduced-form representation of a VAR process can be estimated by OLS. Using the structural innovations will therefore require they first be "recovered" using the relationship between the structural and reduced forms of the VAR process. To make this point explicit, consider the error terms from the reduced form VAR used above, $\hat{e}_t = \mathbf{B}^{-1}\hat{\epsilon}_t$, or:

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} = \begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{bmatrix} \tag{68}$$

Multiplying this out gives:

$$e_{1t} = \frac{\epsilon_{yt} - b_{12}\epsilon_{zt}}{1 - b_{12}b_{21}} \tag{69}$$

$$e_{2t} = \frac{\epsilon_{zt} - b_{21}\epsilon_{yt}}{1 - b_{12}b_{21}} \tag{70}$$

The immediate implication is that one cannot use the reduced form errors to make inferences about cause and effect between endogenous variables in a VAR. This is because movements in either error term can be due to movements in either of the structural shocks. That is, the source of movements in the error terms cannot be identified.

How can the structural shocks be recovered? Because the reduced form representation is a transformation of the structural form, there already exists a relationship between the coefficients and errors in the reduced form with the coefficients and structural shocks in the structural form. Equations (69) and (70) are a good example of this relationship. They specify the error terms (which are known from the estimation) in terms of coefficients and structural shocks from the structural representation. Unfortunately, there are not enough of these equations in the reduced form to fully recover all coefficients and structural shocks from the structural representation. The reason is that there are more unknowns in the structural VAR than estimated coefficients and errors in the reduced form VAR.

To see this, notice that the structural representation, summarized by equations (41) and (42), has ten unknowns: (b_{10}, b_{20}, b_{12}, b_{21}, γ_{11}, γ_{12}, γ_{21}, γ_{22}, σ_y, σ_z). The σ_y and σ_z are standard deviations of y_t and z_t, respectively. Compare this with the reduced form representation, summarized by equations (46) and (47), which yield nine known values (a_{10}, a_{20}, a_{12}, a_{21}, a_{11}, a_{22}, $var(e_{1t})$, $var(e_{2t})$, $cov(e_{1t}, e_{2t})$). The variance of y_t or z_t cannot be used as the tenth known value, as these are both implicit in the variances of the error terms, i.e. they are required to calculate these variances. The reduced form system has only nine known values because the covariances of the different error terms appear twice in the variance-covariance matrix, but are the same. That is, $cov(e_{1t}, e_{2t}) = cov(e_{2t}, e_{1t})$, so only one of these can be used. In order to recover the structural shocks, one of the coefficient estimates in the structural system must be restricted in some way by the modeler. The enormous literature on using VARs for policy analysis boils down to different ways of generating this restriction and applying it to questions of interest.

The explanation on identification to this point has focused on the full VAR representation with two variables. To generalize from the two variable case involves a substantial increase in notation, and it is common to work only in terms of the variance-covariance matrices of error terms ($\mathbf{\Sigma_e}$) and innovations $\mathbf{\Sigma_\epsilon}$, the variances of the structural innovations, and the elements of the matrix $\mathbf{B}$. In addition to simplifying the analysis, working with only these elements puts everything in terms of the items of ultimate interest, which are the structural innovations.

The relationship of interest comes from the reduced form of the VAR: $\hat{e}_t = \mathbf{B}^{-1}\hat{\epsilon}_t$. In estimating the reduced form equation, there is no coefficient matrix on $\hat{e}_t$, but its variance-covariance matrix is

computed. To use this information, take the variance of the errors:

$$var(\hat{e}_t) = E(\hat{e}_t \hat{e}_t) = E(\mathbf{B}^{-1}\hat{\epsilon}_t \hat{\epsilon}_t^{'}(\mathbf{B}^{-1})^{'}) = \mathbf{B}^{-1}E(\hat{\epsilon}_t \hat{\epsilon}_t^{'})(\mathbf{B}^{-1})^{'} \tag{71}$$

This can be simplified further to read:

$$\Sigma_{\hat{e}} = \mathbf{B}^{-1}\mathbf{\Sigma}_{\hat{\epsilon}}(\mathbf{B}^{-1})^{'} \tag{72}$$

This is easier to see when expanded in the n variable case:

$$\begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_1 e_2} & ... & \sigma_{e_1 e_n} \\ \sigma_{e_2 e_1} & \sigma_{e_2}^2 & ... & \sigma_{e_2 e_n} \\ ... & ... & ... & ... \\ \sigma_{e_n e_1} & \sigma_{e_n e_2} & ... & \sigma_{e_n}^2 \end{bmatrix} = \begin{bmatrix} 1 & b_{12} & ... & b_{1n} \\ b_{21} & 1 & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{\epsilon_1}^2 & 0 & ... & 0 \\ 0 & \sigma_{\epsilon_2}^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_{\epsilon_n}^2 \end{bmatrix} \left(\begin{bmatrix} 1 & b_{12} & ... & b_{1n} \\ b_{21} & 1 & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & 1 \end{bmatrix}^{-1} \right)^{'}$$

This form explicitly shows the known and unknown items in the system, assuming the reduced form of the structural equation has already been estimated. This means that the elements of $\Sigma_{\mathbf{e}}$ are known. Because this matrix is symmetric (a square matrix equal to its transpose), it contains $\frac{n^2+n}{2}$ distinct elements. This is calculated by noting that there are n elements along the principal diagonal, $(n-1)$ elements along the first-off diagonal, $(n-2)$ elements along the second-off diagonal, and so on, with two corner elements. This gives a total of $\frac{n^2+n}{2}$ known elements.

The unknowns are the elements of $\mathbf{B}$ and the variances of the structural innovations. The covariances are zero in $\mathbf{\Sigma}_{\hat{\epsilon}}$ because the structural innovations are uncorrelated with each other by definition. $\mathbf{B}$ contains (n^2-n) elements and there are n unknown variances, one for each structural innovation. This gives a total of n^2 unknowns. We have $\frac{n^2+n}{2}$ known elements and n^2 unknowns, meaning that an additional $\frac{n^2-n}{2}$ restrictions must be imposed to recover the structural innovations. In the two-variable example this means there are 3 known elements and 4 unknown elements, so one additional restriction is required.

There is on common modification to this system. It is customary to normalize $\mathbf{\Sigma}_{\hat{\epsilon}}$ to be the identity matrix (i.e. to set the variance of each structural innovation to 1). To account for this change, the main diagonal of the $\mathbf{B}$ matrix can no longer be 1. The system above now reads:

$$\begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e1e2} & ... & \sigma_{e1e_n} \\ \sigma_{e_2e_1} & \sigma_{e_2}^2 & ... & \sigma_{e_2e_n} \\ ... & ... & ... & ... \\ \sigma_{e_ne_1} & \sigma_{e_ne_2} & ... & \sigma_{e_n}^2 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & ... & b_{1n} \\ b_{21} & b_{22} & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & b_{nn} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{bmatrix} \left(\begin{bmatrix} b_{11} & b_{12} & ... & b_{1n} \\ b_{21} & b_{22} & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & b_{nn} \end{bmatrix}^{-1} \right)'$$

Compare this representation with the one above. This normalization does not change the number of known or unknown variables. It is done to allow easy comparison of impulse responses (because each innovation is increased by one standard deviation, now one unit), and for use in computing the Choleski decomposition as described below. Using the normalization in the two-variable case is described in detail in the next section.

The General Identification Solution Procedure

Choosing identifying restrictions for a VAR can be viewed in the context of solving a system of equations. To see why, consider the general n variable case from above, where our interest is in working with the variance-covariance matrices as in equation (71):

$$\Sigma_{\hat{e}} = \mathbf{B}^{-1}\mathbf{\Sigma}_{\hat{\epsilon}}(\mathbf{B}^{-1})'$$

Assume that the variances of the structural innovations have been normalized to one, so that there are coefficients along the diagonal of the **B** matrix as described above. Now $\mathbf{\Sigma}_{\hat{\epsilon}}$ is the same as the identity matrix ($\mathbf{I}_n$) so this can be simplified:

$$\Sigma_{\hat{e}} = \mathbf{B}^{-1}(\mathbf{B}^{-1})' \tag{73}$$

The goal is to find a matrix **B** that solves this system of equations, given that the number of coefficients to be determined in **B** does not exceed the number of equations. As explained above, additional identifying restrictions be need to be imposed to reduce the number of coefficients in order to match the number of equations. The most common method of solving this system, recursive identification, is outlined in the next section.

Recursive Identification

The simplest and most common form for recovering structural shocks (identification) is called recursive identification. The idea is to impose enough restrictions on the system so that there are as many estimated values as unknowns in the underlying structural equations. This is best illustrated in the two-variable case above, with the variance-covariance matrix of the structural shocks the same as

the identity matrix. This means that b_{11} and b_{22} are no longer 1, as was assumed previously. Begin with equations (41) and (42):

$$b_{11}y_t = b_{10} - b_{12}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \epsilon_{1t}$$

$$b_{22}z_t = b_{20} - b_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \epsilon_{2t}$$

It was shown earlier that estimation of the reduced form yields 9 parameter/coefficient values, but the structural system has 10 unknowns. Recursive identification provides this 10th parameter/coefficient value, and amounts to setting one of the coefficient values equal to zero. Here, set b_{12} equal to zero. Applying this restriction alters equations (41) and (42), they now read:

$$b_{11}y_t = b_{10} + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \epsilon_{1t} \tag{74}$$

$$b_{22}z_t = b_{20} - b_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \epsilon_{2t} \tag{75}$$

The $-b_{12}z_t$ falls out of the first equation in this case. In matrix form, we now have:

$$\mathbf{B} = \begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{B}^{-1} = \frac{1}{b_{11}b_{22}} \begin{bmatrix} b_{22} & 0 \\ b_{21} & b_{11} \end{bmatrix}$$

An important point to note is that $\mathbf{B}^{-1}$ is lower triangular.[8] Next, as before use the fact that $\hat{e}_t = \mathbf{B}^{-1}\hat{\epsilon}_t$ and expand:

$$e_{1t} = \frac{1}{b_{11}}\epsilon_{1t} \tag{76}$$

$$e_{2t} = \frac{-b_{21}}{b_{11b_{22}}}\epsilon_{yt} + \frac{1}{b_{22}}\epsilon_{2t} \tag{77}$$

Look closely at these equations. Because b_{12}=0, ϵ_{2t} does not impact y_t in the current period through e_{1t}. Again, the innovation to z_t can no longer change y_t contemporaneously. Notice that y_t still contemporaneously changes z_t. Does this restriction make sense? It depends on the particular variables in the system and the question that is being answered. The main idea is that recursive identification amounts to specifying which of the endogenous variables contemporaneously impact the others.

This is sometimes referred to as ordering the variables. This is because a lower triangular matrix $\mathbf{B}$ provides the order by which the innovations can impact the endogenous variables. If one variable is

[8] A lower triangular matrix is a square matrix where the values above the main diagonal are zero.

ordered before another, it means that the structural innovation to that variable can impact the other variable in the current period. The corollary is that if one endogenous variable is ordered after another, its innovations are unable to change the first in the current period.

Recursive identification always results in a lower triangular matrix $\mathbf{B}$.[9] This fact eases the computation of VARs because the equation $\Sigma_{\hat{e}} = \mathbf{B}^{-1}(\mathbf{B}^{-1})'$ can be solved by using a Choleski decomposition to choose $\mathbf{B}^{-1}$.[10] This is identical to what is done in the two-variable case manually, but can be automated by a software package. The key point to realize is that there is a unique Choleski decomposition for each ordering of the endogenous variables. Again, the ordering will change the solution to the equation above, with each different ordering resulting in a unique $\mathbf{B}^{-1}$.

When is Recursive Identification Appropriate?

Choosing an ordering through recursive identification imposes strong assumptions on the VAR and should be used carefully. Kilian (2011) provides some guidance on when using such an identification method is appropriate. The main point is that such ordering should be done only if it can be justified on economic grounds. This is because the modeler specifies a causal change to identify the system rather than learning about such causal relationships from the data. There should be some reason behind why such a solution is imposed on the shocks for it to be defensible.

What are possible economic grounds for different orderings? One can appeal to economic theory in some cases, although the restrictions will only be as defensible as the theory. In certain cases there may be information delays that can be exploited, ruling out instantaneous feedback. Similarly, there may be physical constraints, such as investment lags. Others can exploit detailed institutional knowledge or assumptions about market structure. In the oil industry this can often refer to OPEC behavior and limitations. Whatever the reason, the ordering chosen will determine the results generated, and so must be justified if this type of identification technique is used.

Examples of VARs Continued

Example 1 Continued: The Price of Oil, Kilian (2009)

In the example of the oil market VAR specified by Kilian (2009), the following error structure is presented:

[9] One can always change the ordering of the equations so that the resulting recursive identification yields a lower triangular matrix.

[10] Roughly, the Choleski decomposition is a decomposition of a matrix ($\Sigma_{\hat{e}}$) into the product of a lower-triangular matrix ($\mathbf{B}^{-1}$) and its transpose [$(\mathbf{B}^{-1})'$].

$$\begin{bmatrix} e_t^{\Delta prod} \\ e_t^{rea} \\ e_t^{rpo} \end{bmatrix} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix}^{-1} \begin{bmatrix} \epsilon_t^{\text{oil supply shock}} \\ \epsilon_t^{\text{aggregate demand shock}} \\ \epsilon_t^{\text{oil specific-demand shock}} \end{bmatrix}$$

First notice that the structural innovations do not specifically correspond to the specific variables on the left-hand side. These innovations cause unexpected movements in the left-hand side variables, but can be interpreted more broadly than just movements in those variables. This system is ordered recursively. The first assumption is that the change in oil production cannot be altered by economic activity or the oil price in the current period, i.e. it is ordered first. This is another way of saying that the oil supply curve is inelastic within the current month (the data in this study are monthly). Kilian (2009) uses various evidence to support this assumption. Similarly, the oil price does not change economic activity in the current period, meaning that economic activity is ordered second, another assumption which is also justified using different evidence. Given these identifying (or ordering) assumptions, the underlying structural innovations can be identified, and Kilian (2009) uses innovation accounting to present his results.

Example 2 Continued: The Impact of Monetary Policy, Cochrane (1994)

In the example of the impact of monetary policy and associated VARs by Cochrane (1994), the following error structure for the simplest case is presented:

$$\begin{bmatrix} e_t^{m2} \\ e_t^{y} \\ e_t^{p} \end{bmatrix} = \begin{bmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{bmatrix}^{-1} \begin{bmatrix} \epsilon_t^{\text{money supply}} \\ \epsilon_t^{\text{gdp}} \\ \epsilon_t^{\text{price level}} \end{bmatrix}$$

The ordering imposed in this case is strong as well. Neither GDP nor the price level can change the money supply in the current period (a quarter), and the price level cannot alter GDP in the current period either. However, changes in the money supply can instantaneously change both GDP and the price level. This is one of many variations that Cochrane (1994) uses in trying to isolate the impact of monetary policy on the macro-economy. This set-up is used to identify the system, and innovation accounting results are then presented.

Innovation Accounting

There are no substantial changes to the procedures outlined in the sections on impulse responses, variance decompositions, or historical decompositions above. The interpretation of these tools must now consider that each is dependent on the ordering of the variables (or whichever identification technique is used). A different ordering will give different impulse response functions and will change the relative importance of the structural innovations in either decomposition method. As a final note on the impulse response analysis, the variables which are ordered first will not respond instantaneously

to the innovations in later variables. The plots will reflect a response in the period following the shock due to this ordering choice.

Recursive Identification Using Bayesian Methods

There are two ways in which recursive identification can be used in combination with Bayesian methods. The first, which is not discussed in detail here, is to use a prior distribution over the recursive identification scheme itself (the ordering of the **B** matrix). One can then back out the posterior distribution of this ordering and use that for innovation accounting. The second is to use the posterior distribution of the reduced form coefficients for innovation accounting. In this case, the same procedures used above still apply, as do the equations summarizing the resulting impulse responses and variance decompositions. There is no need to alter any identification procedures if using recursive identification with Bayesian methods.

The changes when using Bayesian estimation in this manner stem from the fact that the reduced form coefficient values are no longer point estimates, but are associated with a distribution. Take for example the impulse response functions, and assume that the reduced form model has been estimated, and the corresponding posterior distribution of the estimates is available. It was shown above that the impulse response functions could be written (in the two variable case) as:

$$\begin{bmatrix} y_t \\ z_t \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} + \sum_{i=0}^{\infty} \begin{bmatrix} \phi_{11}(i) & \phi_{12}(i) \\ \phi_{21}(i) & \phi_{22}(i) \end{bmatrix} \begin{bmatrix} \epsilon_{1t-i} \\ \epsilon_{2t-i} \end{bmatrix}$$

Recall that the $\phi(i)$ coefficients, called impact multipliers, summarize the impacts of changes in the structural innovations on the endogenous variables. For example, $\phi_{11}(0)$ is the instantaneous impact of ϵ_{1t} on y_t, and $\phi_{12}(0)$ is the instantaneous impact of ϵ_{2t} on y_t. The ϵ are the structural innovations, and the x and y are the endogenous variables. One option is to use the mean of the posterior distribution to get point estimates for the coefficients, use these in combination with the identifying assumptions to solve for the ϕ, and then generate the impulse responses using the equation above.

An alternative is to simulate the distribution of impulse responses. In this case one can sample (arbitrarily pick points) from the posterior distribution of the reduced form coefficients, use these as point estimates in combination with the identifying assumptions, and generate the impulse response. Continually sampling over different values of the posterior distribution will lead to a variety of different of impulse responses. It is then standard to use the mean response from these impulse responses as summary, and confidence intervals can be generated as described above.

The procedure is the same for variance decompositions and historical decompositions. Once the posterior distributions of the reduced form coefficients are found, one can use their mean as a point estimate in combination with the recursive identification to generate point estimates for the variance

or historical decompositions. The alternative is to sample from the posterior distribution of the reduced form coefficients, use a sample from this posterior in combination with the recursive identification scheme to generate a distribution of values for either distribution. A mean of these can then be used as a summary statistic.

Strengths and Weaknesses of Policy Analysis With VARs

VARs are widely used for policy analysis, and there are pros and cons to this approach. A very appealing aspect of using VARs for policy analysis is the flexibility of the procedure. There is substantial freedom for the modeler to add or drop variables, add or drop lags, and vary time periods as required. The technique also has wide policy applicability, and has been used in contexts relating to monetary policy, fiscal policy, economic uncertainty, factors impacting the oil price, and many others. VARs are often the only econometric technique available which can incorporate many variables without specifying them to be exogenous or imposing a strict model structure. Impulse responses, variance decompositions, and historical decompositions are also very powerful tools for analysis.

The drawbacks of using VARs for policy analysis boil down to identifying restrictions. Often these are ad-hoc and may be hard to justify based on any type of theory. This is particularly true of the recursive identification explained in this section, but is also true of alternative techniques. Where economic theory can be used, as with the sign-restricted VARs outlined below, there is often a problem of finding a unique solution, or at least the most likely solution. Another objection to using VARs for policy analysis is the imposition of a linear structure combined with the assumption that structural innovations exist. Can the movements really be interpreted as structural innovations?

Forecasting With VARs

Forecasting with VARs can be simpler than conducting policy analysis because the simplest case requires only the estimated coefficients from a reduced-form VAR. As with any empirical approach, complications arise in choosing the variables to include, as well as the number of lags to use, and these are addressed in the first section below. The second section takes up the fact that estimated VARs are often overparameterized, and discusses some options for dealing with this problem. The section concludes with some of the pros and cons of using VARs for forecasting based on Carnot et al. (2011).

General Overview

A VAR can be used for forecasting once its coefficients have been estimated. The coefficient estimates are based on data through some time period, T. Once these estimates are available, forecasts can be generated for an arbitrary number of periods ahead. A model following Enders (2010) makes this clearer. Suppose a reduced-form VAR model is estimated on data up until period T:

$$\hat{x}_T = \mathbf{A}_0 + \mathbf{A}_1 \hat{x}_{T-1} + \hat{e}_T$$

Where $\hat{x}_T$ and $\hat{x}_{T-1}$ are $kx1$ vectors of endogenous variables, the $\mathbf{A}$ are kxk matrices, and $\hat{e}_T$ is a $kx1$ vector of errors. The values of $\mathbf{A}_0$ and $\mathbf{A}_1$ are known, and the one-step ahead forecast is given by:

$$E_T \hat{x}_{T+1} = \mathbf{A}_0 + \mathbf{A}_1 \hat{x}_T$$

In this equation E_T is the conditional expectation of x_{T+1} given information available at time T. Because of the expectation there is no error term in this forecast. A two-step ahead forecast is generated in a similar fashion:

$$E_T \hat{x}_{T+2} = \mathbf{A}_0 + \mathbf{A}_1 \hat{x}_{T+1} = \mathbf{A}_0 + \mathbf{A}_1(\mathbf{A}_0 + \mathbf{A}_1 \hat{x}_T)$$

In this manner n-step ahead forecasts at time T can be generated from the coefficient estimates and values of the variables at T.

In using this basic procedure, the stationarity of the data and overparameterization of the VAR become more important than when using a VAR for policy analysis. This is because the forecasts are based on estimated coefficients, and if these values cannot be trusted, either because of non-stationary data or because they are statistically insignificant, the resulting forecasts may be unreliable. Additionally, even if the series is stationary over the sample period, the relationships between the variables may not be stable over time. The next two sections look closely at these issues.

Stationarity of Data

The importance of coefficient estimates for forecasting with VARs requires addressing the stationarity characteristics of the data. Because basic OLS assumptions are violated when estimating with non-stationary data, the coefficient estimates in this case do not provide useful information. At this point a forecaster can either transform the data (i.e. difference, de-trend, or filter) or use an error correction model if there is non-stationary data which is cointegrated.

The different procedures for transforming the data can lead to the loss of important characteristics of the time series, and are avoided if at all possible. But some transformation often must be done to use the coefficient values. One exception where non-stationary data may not need to be transformed is in the presence of a cointegrating relationship. In this case a vector error correction model (VECM) can be used in place of a VAR.

An error correction model is designed to incorporate long-run information available from the cointegration of variables, although it allows for the use of both transformed and non-transformed

data. This is highlighted using an example from Kennedy (2008) for the univariate case:

$$y_t = B_0 + B_1 x_t + B_2 x_{t-1} + B_3 y_{t-1} + e_t \tag{78}$$

Suppose that both x and y are individually non-stationary and economic theory suggests that in the long-run they are cointegrated, with relationship: $y = \psi + \theta x$. The equation above can be manipulated using this relationship to yield an error correction model:

$$\Delta y_t = B_1 \Delta x_t + (B_3 - 1)(y_{t-1} - \psi - \theta x_{t-1}) + e_t \tag{79}$$

The second term on the right-hand side is the error correction term. Because this relationship is believed to hold in the long-run, it adjusts for any errors as either variable deviates from the long-term relationship. For example, if y grows too quickly so does the last term, and because the last coefficient is negative ($B_3 < 1$ due to stationarity), Δy_t is reduced in recognition of this error. The strength of the ECM model is that is uses both differenced variables as well as levels values. In theory this should give it an advantage in generating coefficient estimates.

This ECM can be extended to a vector error correction model (VECM) in the same manner univariate regressions are extended to VARs. In the presence of cointegrated variables, a VECM can be used to generate forecasts rather than a VAR with transformed data. The procedure for generating n-step ahead forecasts is as described above.

Near-VARs and Bayesian VARs

The overparameterization issue is usually overcome by changing the type of VAR used. One approach is to use near-VARs, in which the original VAR is "shrunk" to get rid of insignificant coefficient values. The other approach is to use Bayesian methods and specify a prior distribution on each coefficient, which can be combined with the data to yield a posterior distribution for each coefficient. These two methods are discussed in turn.

Near-VARs

A common approach to overcome overparameterization of a VAR is to throw out the insignificant coefficients and re-estimate the VAR. This might mean discarding variables entirely, or just certain lags of the variables, and is termed a near-VAR. One complication with this procedure is that the right-hand sides of the equations in the VAR are no longer the same when estimating. Because the errors are correlated across equations in the reduced-form, OLS is no longer efficient. The most common way to avoid this complication is by using seemingly unrelated regression estimation (SURE).

The SURE method is used because the reduced form equations are not a set of simultaneous

equations (all endogenous variables on the right-hand side are predetermined), but rather because they are related only through error terms. SURE overcomes this problem by writing the set of individual equations in a VAR as one giant equation. Now the correlations of error terms are captured by the variance-covariance matrix of the error vector in this large regression. The covariances in this matrix can be deduced from the correlations between residuals in the separate equations. The large equation is then estimated by generalized least squares (GLS), which minimizes a weighted sum of square residuals, and the variance-covariance matrix provides the weights. Once this estimation is complete, the coefficient values can be used for forecasting as above.

The other issue with using a near-VAR is deciding which coefficient estimates to drop. Dropping these values works on the assumption that every insignificant coefficient estimate is actually zero, which may not be the case. Forecaster judgement can be used to drop only the parameter estimates which are not of interest, or an iterative procedure can be adopted whereby different combinations of coefficients are dropped, and the one giving the best in-sample accuracy is chosen. One can think of this as varying the lag lengths in each equation, and this lag length could possibly be different for each variable. The issue with this approach is the large number of variations which could be chosen, and this effectively makes this procedure unworkable in larger VARs. An alternative is to use Bayesian VARs, which are discussed next.

$Bayesian\ VARs$

Bayesian VARs are an alternative to near-VARs that do not require the forecaster to take an all-or-nothing approach on the values of coefficients. That is, dropping a variable or a lag of a variable effectively sets its value to zero. This can be avoided with Bayesian methods by placing a prior probability on the value of each coefficient. In this way, fuzzy restrictions can be placed on the coefficients, which may be over-ridden by the data.

As with near-VARs, specifying a prior distribution on each coefficient becomes unworkable very quickly, and there are several approaches that can be taken to simplify the problem. The most popular is to specify the so-called Minnesota prior for the model coefficients, see Ciccarelli and Rebucci (2003) for a full exposition of this technique. When using this prior, the variance of the error term in each equation is assumed fixed and known. This reduces the dimensionality of the problem, as the prior distribution is no longer a joint distribution of the coefficients and error variances, but of just the coefficients themselves. Assuming this is normally distributed leads to:

$$p(\beta_g) = N(\bar{\beta}_g, \bar{\Omega}_g) \tag{80}$$

Where $\bar{\beta}_g$ and $\bar{\Omega}_g$ refer to the prior mean and variance-covariance matrix of β_g, and the subscript g is used to refer to the fact that these come from the gth equation of the VAR. As mentioned before, the

variance-covariance matrix of the residuals, Σ, is assumed to be known and fixed. To reiterate, with the Minnesota prior we assume that the variance-covariance matrix of reduced form errors is fixed and known and that the prior distribution of the coefficient estimates in each equation is normally distributed.

Assuming that the reduced form errors are also normally distributed, the likelihood function for the VAR (and each equation within the VAR) can be derived. The posterior distribution for the coefficients in each equation can be written:

$$p(\beta_g|Y) = p(\bar{\beta}_g)L(Y|\bar{\beta}_g, \sigma_{gg}^2) \tag{81}$$

Where $L(Y|\bar{\beta}_g, \bar{\Omega}_g)$ is the likelihood function, and the posterior is conditional on the data, Y, and σ_{gg}^2 is the variance of reduced form errors. After some messy manipulations of the likelihood function, the posterior distribution of the coefficients in each equation can be shown to be normally distributed, or:

$$p(\beta_g|Y) = N(\tilde{\beta}_g, \tilde{\Omega}_g) \tag{82}$$

And $\tilde{\beta}_g$ depends on the data, β_g, $\bar{\Omega}_g$, and σ_{gg}^2. The other parameter, $\tilde{\Omega}_g$, depends on the data, $\bar{\Omega}_g$, and σ_{gg}^2. Each of these is known, meaning that this is a closed form solution for the posterior distribution. Notice also that this is only for equation g, but neither the prior distribution nor the posterior distribution of β_g depend on any other equations in the VAR. There is prior and posterior independence between equations. The posterior distribution of the coefficients for each equation can be generated by selecting values for β_g and plotting the resultant probability.

The issue of how to specify the parameter values of interest is an important one when using a Minnesota prior. The variance-covariance matrix of the residual errors, Σ, is diagonal with variances σ_{gg}^2 on the diagonal. One way to estimate these variances is to specify an $AR(p)$ model for each variable g and use the variance of the residuals. The prior mean ($\bar{\beta}_g$) and variance-covariance matrix ($\bar{\Omega}_g$) are also unknown, but are specified through hyper-parameters. Using hyper-parameters to summarize the multidimensional normal distribution on the prior distribution of the coefficient vector can greatly simplify the computations and reduce choices which must be made by the modeler (Koop, 2003).

The original Minnesota prior specified six hyper-parameters, and these values were based on some observations of time series processes. The first is that most macroeconomic time series are well-represented by random walk processes. Thus the mean coefficient vector specifies a mean of 1 for the first lag of variable g in equation g and zero for all others. The second is that the variance of lags is less important than the current variance. The other five hyper-parameters control the specification of standard deviations for lagged coefficient values of both the variable under consideration and other

variables. In total, using a Minnesota prior can yield a posterior distribution with specification of only six hyper-parameters (irrespective of how many variables or lags are in the VAR) and the variances of the residual errors. Alternatives to the Minnesota prior which move away from some of its assumptions include the diffuse prior and conjugate prior distributions. See Ciccarelli and Rebucci (2003) for more details on these.

Once the posterior distribution is generated there are two methods for generating point forecasts. The first is to use the mean of the posterior distribution as a point estimate for the coefficients, and then to forecast as described above in the general case. The second is to generate this point forecast by averaging over the forecasting function by using the posterior distribution of the coefficients as weights. This differs from the first method because it uses more than the mean of the posterior, it uses the entire distribution to generate the weighted average. If the density forecast is desired, as opposed to a point forecast, one can integrate the conditional distribution of the forecast with the posterior distribution of the coefficients.

Strengths and Weaknesses of Forecasting With VARs

VAR models have been shown to be as good as other methods in generating forecasts. They are smaller than macroeconometric models, and easier to estimate. There is also no need to make assumptions about exogenous variables when forecasting with a VAR. VAR models have been used to help sort survey-based information, can also be used to identify causal relationships (this requires some identifying assumptions discussed above), and may help to interpret ongoing developments through their use in counterfactual simulations.

However, forecasting using VARs still requires specifying which variables to include in the estimations, as well as how many lags to include. VARs are also larger than comparable ARMA models. They may also be difficult to interpret in economic terms, which is the strength of using a macroeconometric model.

Other Identification Techniques

Recursive identification has been criticized on many grounds, but particularly because it may be difficult to incorporate relevant economic theory. Structural, long-run, and sign-restricted identification techniques all try to overcome this limitation, and are discussed below.

Structural Restrictions

Structural restrictions use economic theory in specifying the identifying restrictions among structural shocks in a VAR. While this can be done in a recursive set-up as well, structural VARs move away from lower-triangular coefficient matrices. The technique can be illustrated following Enders (2010), beginning with:

$$\begin{bmatrix} \sigma^2_{\hat{e}_1} & \sigma_{e_1e_2} & ... & \sigma_{e_1e_n} \\ \sigma_{e_2e_1} & \sigma^2_{\hat{e}_2} & ... & \sigma_{e_2e_n} \\ ... & ... & ... & ... \\ \sigma_{e_ne_1} & \sigma_{e_ne_2} & ... & \sigma^2_{\hat{e}_n} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & ... & b_{1n} \\ b_{21} & b_{22} & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & b_{nn} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{bmatrix} \left(\begin{bmatrix} b_{11} & b_{12} & ... & b_{1n} \\ b_{21} & b_{22} & ... & b_{2n} \\ ... & ... & ... & ... \\ b_{n1} & b_{n2} & ... & b_{nn} \end{bmatrix}^{-1} \right)'$$

Or more succinctly:

$$\Sigma_{\hat{e}} = \mathbf{B}^{-1}(\mathbf{B}^{-1})' \tag{83}$$

The goal is to find a matrix **B** that solves this system of equations, given that the number of coefficients to be determined in **B** does not exceed the number of equations. As explained above, additional identifying restrictions need to be imposed to reduce the number of coefficients in order to match the number of equations. A recursive scheme restricts **B** to be lower triangular. One might also choose the value of coefficient(s) in this matrix, or even restrict a variance to a certain value. These latter two are done infrequently, as such values are rarely known beforehand. Another approach sets certain values of **B** to zero, but does not specify that it must be lower triangular, and this is the structural decomposition discussed here.

The benefit of this approach is that it is easier to inform with economic theory. A three variable example helps to make this clear, as before $\hat{e}_t = \mathbf{B}^{-1}\hat{\epsilon}_t$. Define $\mathbf{C} = \mathbf{B}^{-1}$ to ease notation. In the Choleski case, this system yields three equations:

$$e_{1t} = \epsilon_{1t}$$

$$e_{2t} = c_{21}\epsilon_{1t} + \epsilon_{2t}$$

$$e_{3t} = c_{31}\epsilon_{1t} + c_{32}\epsilon_{2t} + \epsilon_{3t}$$

Notice that c_{12}, c_{13}, c_{23} are all set to zero because the matrix is lower-triangular. This system can be solved because there are three equations and three unknowns. However, one might also choose not to set those three coefficients to zero and have:

$$e_{1t} = \epsilon_{1t} + c_{13}\epsilon_{3t}$$

$$e_{2t} = c_{21}\epsilon_{1t} + \epsilon_{2t}$$

$$e_{3t} = c_{31}\epsilon_{2t} + \epsilon_{3t}$$

In this case c_{32} has been set to zero and c_{13} is no longer zero. The logic behind such an ordering may come from economic theory and depends on the specific variables. The point is that recursive ordering is not the only possible way to identify the system.

The standard way to solve these problems is to first use OLS to find the reduced form coefficient matrices. Next, one uses maximum likelihood techniques to solve $\Sigma_{\hat{e}} = \mathbf{B}^{-1}(\mathbf{B}^{-1})'$, with the usual assumption that the errors are normally distributed. In the Choleski case this second step does not need to be taken because it is already known that the **B** matrix is lower triangular.

While this might seem like a standard MLE problem, it is more complicated. Once the likelihood function has been derived using the distribution of the errors, one usually substitutes in for the errors with the assumed model. In this case there is no model to substitute in, as there is no data that is used, all we have are the restrictions on the variance/covariance matrix. Because of this, there is no way to verify that the solution to the MLE estimation is unique. In this context this uniqueness is sometimes referred to as identification. Local identification yields matrices which are unique with values close to the starting guesses for MLE, while global identification does so irrespective of where the original guesses originate. The takeaway from this discussion is that simply counting the number of knowns and unknowns and imposing the difference in restrictions does not guarantee a unique solution unless the Choleski decomposition is used.

Long-run Restrictions

Long-run restrictions are an alternative means of identifying structural shocks. These move the focus from the instantaneous and continuing period-by-period impacts given by structural restrictions to those based on the total impact. To use the method, at least one of the variables must be I(1). The two variable example given in Enders (2010) illustrates this method well. Begin with the $VMA(\infty)$ representation of a VAR:

$$\Delta y_t = \Sigma_{k=0}^{\infty} c_{11}(k)\epsilon_{1t-k} + \Sigma_{k=0}^{\infty} c_{12}(k)\epsilon_{2t-k} \tag{84}$$

$$z_t = \Sigma_{k=0}^{\infty} c_{21}(k)\epsilon_{1t-k} + \Sigma_{k=0}^{\infty} c_{22}(k)\epsilon_{2t-k} \tag{85}$$

The first variable is in differenced form because it is I(1). Recall that the coefficients $c_{ab}(k)$ summarize the impact of the structural shocks k periods ago in the equation relating variable b to variable a. So $c_{12}(5)$ in the equation above represents the impact of the structural shock to z_{t-5} on Δy_t. More compactly:

$$\begin{bmatrix} \Delta y_t \\ z_t \end{bmatrix} = \sum_{k=0}^{\infty} \begin{bmatrix} c_{11}(i) & c_{12}(i) \\ c_{21}(i) & c_{22}(i) \end{bmatrix} \begin{bmatrix} \epsilon_{1t-k} \\ \epsilon_{2t-k} \end{bmatrix} \tag{86}$$

The basic idea is that many variables can be decomposed into temporary and permanent components, and one can use this information to identify the structural innovations from an estimated VAR. This requires disassociating the structural shocks with particular variables, and thinking of the structural shocks as exogenous variables which impact the endogenous variables in the system. Because y is non-stationary it has both a permanent and temporary component, and the identification procedure is to assume that one shock has a temporary effect on the y sequence, and the other a permanent effect.

Suppose that the y sequence is GDP and the shock associated with this series, ϵ_1, is an aggregate demand shock. A plausible long-run restriction is that aggregate demand shocks only have temporary effects on GDP, but not permanent. In the context of the VAR, one could say that the cumulative effect of this shock on GDP is zero, or:

$$\Sigma_{k=0}^{\infty} c_{11}(k)\epsilon_{1t-k} = \quad 0 \quad = \Sigma_{k=0}^{\infty} c_{11}(k) \tag{87}$$

Assuming that all of the variables are stationary (this is a key assumption for imposing long-run restrictions), some tedious manipulations will allow the system to be identified. The main restriction imposed in this type of system is to specify the temporary versus permanent impact on the non-stationary variable(s) in the system.

Sign-Restricted Identification

Sign-restricted VARs are reviewed in this section following Fry and Pagan (2007) and Kilian (2011). This identification technique is considered an alternative to recursive, structural, or long-run methods. The section begins with some motivation on the approach in general, outlines the basic technique, and follows with a discussion of the appropriateness of using sign-restricted identification.

Motivation

Recursive, structural, and long-run identification techniques all place strong restrictions on the relationships between variables. Both recursive and structural identification (sometimes called short-run restrictions) assume that within the specified period some variables do not respond to movements in other variables. This relationship holds throughout the entire sample, and there is no way to incorporate additional information about magnitudes or directions of the relationships between variables. It is often the case that there is no theory which can justify such restrictions. Similarly, long-run restrictions specify a long-run relationship between at least two of the variables in the VAR over the entire sample period. They also do not allow for incorporation of qualitative or quantitative information on the responses of variables to innovations in other variables.

Sign-restricted VARs have become popular because they overcome these limitations. Instead of excluding variables from responding to innovations in other variables for a specified time period, VARs

identified by sign-restrictions specify the sign of such responses. This is a very powerful advance because it allows more explicit use of theory in identifying the VAR. Take for example the VAR used by Cochrane (1994). It is identified by assuming that within a quarter, innovations in GDP or the price level cannot alter the money supply. But surely the Federal Reserve can respond to price level movements within a quarter by adjusting the money supply. In almost any application of recursive identification to VARs such a situation arises. A sign-restricted VAR instead might specify that the response of the money supply to a positive innovation in the price level is negative. Notice this type of sign-restriction is easier to justify based on theory, but it is also weaker because it does not exclude a response of the money supply to the price level innovation in the current quarter. This is the tradeoff inherent in using sign-restricted VARs, as is explained in detail below.

Basics

The basic idea behind sign-restrictions is similar to that outlined in the recursive case. Begin with equation (83):

$$\Sigma_{\hat{e}} = \mathbf{B}^{-1}(\mathbf{B}^{-1})'$$

Recall that the goal here is to find a matrix $\mathbf{B}$ that solves this system of equations. The Choleski decomposition is one method of solving the system, and this results in a lower triangular matrix. But is this the only solution? No, the solution will change (i.e. the matrix $\mathbf{B}$) if the recursive ordering is changed. Any solution to this equation is not unique, but depends on the ordering, although the Choleski decomposition results in a unique lower-triangular matrix for a particular ordering. This can also be further modified. Consider a different matrix $\mathbf{P}$, where $\mathbf{P}(\mathbf{P})' = \mathbf{I}$. Assume that $\mathbf{B}$ solves equation (83) and rewrite:

$$\Sigma_{\hat{e}} = \mathbf{B}^{-1}\mathbf{P}(\mathbf{P})'(\mathbf{B}^{-1})' \tag{88}$$

This modification still solves the system of equations. However, in most cases $\mathbf{B}^{-1}\mathbf{P}$ is not the same as $\mathbf{B}^{-1}$. Furthermore, it is very unlikely that $\mathbf{B}^{-1}\mathbf{P}$ is lower triangular. This means that along with the recursive solution to the equation embodied in $\mathbf{B}^{-1}$, there are potentially many other solutions to the same model which can be generated by multiplying it by some $\mathbf{P}$.

This creates a problem and an opportunity. Define $\mathbf{D} = \mathbf{B}^{-1}\mathbf{P}$. This equation has as many solutions as there are orthogonal $\mathbf{P}$ matrices, i.e. matrices with the property that $\mathbf{P}(\mathbf{P})' = \mathbf{I}$. How to narrow down the possibilities? One method is to specify sign-restrictions on the responses of the variables. The specified response of every variable to every innovation is not required. We only need to specify the same number as would be required in the recursive identification case. The basic idea is to

randomly draw an orthogonal matrix $\mathbf{P}$ and multiply it by $\mathbf{B}^{-1}$ to find $\mathbf{D}$.[11] This solves equation (83), and allows for simulation of impulse response functions. One can then check each of the impulse responses against the specified sign-restrictions. If all of the sign-restrictions are met, keep the matrix $\mathbf{P}$, otherwise discard. Repeat this procedure many times by continually drawing random $\mathbf{P}$ matrices.

The result of these repeated simulations will be many candidate solutions, all of which match the pre-specified sign restrictions. Although they match these sign-restrictions, they are not the same. The responses which were not pre-specified (presumably including the response of interest) will generally differ. The issue now is how to choose a particular candidate solution or summarize the responses of all candidate solutions. If only qualitative information is required (i.e. how does the price level respond to an increase in money supply?) then one can look at the fraction of the candidate solutions that show a positive versus negative response. Most of the time, however, quantitative information is also desired. There are several ways to get quantitative results.

An early approach was to choose the candidate solution most favorable to the hypothesis of interest. This approach may be useful for some purposes, but most applications require specifying which of the candidates is more likely, which this approach cannot address. Another method has been to only derive confidence intervals without using the point-wise impulse responses. This is feasible, but turns out to be quantitatively demanding.

The most popular method has been to rely on Bayesian methods. In this case the posterior distribution of impulse response functions is calculated. Recall that this is the distribution of the various impulse responses given the data. From here it is unclear how to proceed. Many authors report the median impulse response, as a proxy for a central tendency. The problem with this approach is that the posterior distribution is based on different models. Each random draw of the $\mathbf{P}$ matrix corresponds to a different model. In this case reporting the median does not provide any information that can be used as a measure of central tendency. Some authors have resorted to identifying the most likely model of all the models which generate the posterior distribution. This is a recent advance and computationally demanding.

A different way to approach this problem is to further restrict the VAR. Some specify that cross-correlations between variables must meet certain sign-restrictions. Others use knowledge of a particular market to impose these restrictions. In the oil market, for example, one could add the restriction that oil supply cannot quickly respond to oil price changes (i.e. restrict the price elasticity of oil supply). This is the approach taken in Kilian and Murphy (2011).

[11] Randomly drawing an orthogonal matrix can be complicated, see Fry and Pagan (2007) for details.

Sign-Restricted Identification Using Bayesian Methods

The procedure for sign-restricted identification using Bayesian methods is identical to that using classical methods. The differences show up when generating impulse response functions and variance decompositions. The standard procedure is to generate a posterior distribution for the reduced form coefficients using an inverse Wishart-Gaussian prior on these coefficients and a uniform prior on the **P** matrices (see Kilian and Murphy (2011) for more details).

As with recursive identification, one can then use point estimates from the posterior or sample from the posterior to generate the impulse responses. When using point estimates, one can use the mean of the posterior distribution and combine these with the identifying restrictions to generate impulse responses. Each impulse response corresponds to a different model because the **P** matrix used will be different. One then has the same problem as above in deciding which of the impulse responses to use as a summary response.

The alternative is to sample from the posterior distribution of the reduced form coefficients and combine with the identifying assumptions to generate a distribution of impulse responses. At this point the mean of this impulse response distribution could be used as a summary statistic for each unique **P**. But each of these matrices will have its own impulse response function, and the summary problem exists in this case as well.

When is Sign-Restricted Identification Appropriate?

This approach is most appropriate when other methods of identification cannot be justified. Often this will be in cases where theory clearly shows how variables should react to movements in other variables, but has nothing to say about time frames. In addition, this approach may work better with lower frequency data, as it is difficult to rule out responses over longer periods such as quarters or years.

References

Carnot, Nicolas, Vincent Koen, and Bruno Tissot, *Economic Forecasting and Policy*, 2nd ed., Palgrave Macmillan, 2011.

Ciccarelli, Matteo and Alessandro Rebucci, "Bayesian VARs: A Survey of the Recent Literature with an Application to the European Monetary System," Working Paper WP/03/102, IMF 2003.

Cochrane, John H., "Shocks," *Carnegie-Rochester Conference Series on Public Policy*, 1994, *41*, 295–364.

_ , "Time Series for Macroeconomics and Finance," Lecture Notes 2005. Available at: `http://faculty.chicagobooth.edu/john.cochrane/research/papers/time_series_book.pdf`.

Enders, Walter, *Applied Econometric Time Series*, 3rd ed., Wiley, 2010.

Fry, Renee and Adrian Pagan, "Some Issues in Using Sign Restrictions for Identifying Structural VARs," NCER Working Paper 14, NCER 2007.

Kennedy, Peter, *A Guide to Econometrics*, 6th ed., Wiley-Blackwell, 2008.

Kilian, Lutz, "Not All Oil Price Shocks are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market," *American Economic Review*, 2009, *99*, 1053–1069.

_ , "Structural Vector Autoregressions," CEPR Discussion Paper 8515, CEPR 2011.

_ **and Daniel P. Murphy**, "The Role of Inventories and Speculative Trading in the Global Market for Crude Oil," Mimeo 2011.

Koop, Gary, *Bayesian Econometrics*, 1st ed., Wiley, 2003.

Lutkepohl, Helmut, *New Introduction to Multiple Time Series Analysis*, 1st ed., Springer, 2007.

Nelson, Charles R. and Charles I. Plosser, "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, 1982, *10* (2), 139–162.

Appendix C: The Mechanics of Macroeconometric Models

Overview

This document summarizes some technical details of macroeconometric models for use in both policy analysis and forecasting.[1] To make the document as self-contained as possible, the first section briefly provides a short introduction to basic time series concepts, models, and estimation. The second section gives a general description of the specification, estimation, solution, and testing of macroeconometric models. The final two sections outline the short and long-run theory commonly used in constructing and specifying macroeconometric models.

Preliminary Econometrics

Macroeconometric models, as the name implies, are based on both macroeconomic theory and econometrics. This section outlines some basic econometrics that are useful in understanding the equations found in this class of models. It begins by reviewing time series concepts that are fundamental to macroeconometric models. The second section covers basic time series models, including autoregressive moving average (ARMA) models, vector autoregressions (VARs), and error-correction models (ECMs). The final section gives an overview of different techniques such as ordinary least squares (OLS) and 2-stage least squares (2SLS) which are used to estimate macroeconometric models.

Basic Time Series Concepts

This section covers a variety of basic issues in time series analysis ranging from notational definitions to commonly-used concepts. There are also discussions of testing for unit roots, cointegration, and testing for cointegration because of their importance in macroeconometric modeling.

[1] None of the material covered here is original. As much as possible, the original sources of the equations, explanations, and examples have been cited.

Basic Concepts

A difference equation expresses the value of a variable as a function of its own lagged values, other variables, and time. The equation becomes stochastic if any of the other variables are random, or if random error or disturbance terms are added. The implications of the efficient market hypothesis for stock prices are consistent with a well-known example of a stochastic difference equation, the random walk:

$$y_t = y_{t-1} + \epsilon_t \tag{1}$$

In this case y_t is interpreted as the price of a share on day t and ϵ_t as a random disturbance term with mean zero.

The stochastic error, often called a disturbance or innovation, is very important in time series econometrics, particularly in autoregression analysis. The properties of this process will often be restricted so that it is white noise. A white noise process has zero mean, finite variance, and is serially uncorrelated. The fact that it has zero mean says that over repeated samples the average value of the process will be zero.[2] Looking at equation (1), this means that the stock price today on average is the same as the stock price yesterday.

Recall that the variance of a process measures how far on average it deviates from the mean. It is the expected value of the squared deviation of a sample draw from its mean.[3] The variance is often assumed to be constant over time, or homoscedastic. Serial correlation, also known as autocorrelation, is the correlation between values of the process at different points in time. Correlation, which is based on covariance, measures how much two random variables move together, and ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).[4]

A common additional restriction is to assume that the realizations of the white noise process are distributed normally, making it a Gaussian white noise process. The purpose of assuming that error terms are restricted to be Gaussian white noise is to ease estimation by ordinary least squares (OLS) or maximum likelihood (ML). Not only can the estimation techniques be applied in a straightforward manner, but the resulting standard errors can be used for hypothesis testing (assuming the other assumptions of either technique are met).

The concept of Granger causality is also important in time series analysis. Roughly, one can say that a

[2]The expected value of a process of discrete random variables $\{X_t\}$ is $E[X] = x_1p_1 + x_2p_2 + ... + x_kp_k$, where the x_j are any values that X_t can take and the p_i are their associated probabilities.

[3]Let $E[X] = \mu$, then the variance of $\{X_t\}$ at any point in time t is $Var[X_t] = E[(X_t - \mu_t)^2]$. This is often denoted σ^2 if it is constant over time, or σ_t^2 if it varies. The standard deviation, σ is the square root of the variance.

[4]The autocovariance of $\{X_t\}$ over one period is given by $Cov[X_t, X_{t-1}] = E[(X_t - \mu_t)(X_{t-1} - \mu_{t-1})]$. The autocorrelation is a normalization of this number to lie between -1 and 1, and is computed by dividing the autocovariance by $\sigma_t\sigma_{t-1}$, or the product of the standard deviations of the process at the different times.

variable Granger causes another if an unexpected movement in the first variable helps to forecast the other variable. This is not causality in the usual sense because there might be other variables which affect both of those being tested. There are several ways to test for Granger causality, and these are outlined in Cochrane (2005).

It is also useful to differentiate between structural and reduced-form equations. A structural equation is one which expresses an endogenous variable y_t as being dependent on the current realization of another endogenous variable x_t, and possibly its own lags, the lags of other endogenous variables, current and past values of exogenous variables, and disturbance terms. A reduced-form equation is similar, but any included endogenous variables must be lagged. A reduced form equation does not express one endogenous variable in terms of the current value of another endogenous variable.

Summarizing the variance of a vector autoregressive process requires introduction of the variance-covariance matrix. A variance-covariance matrix, usually denoted $\mathbf{\Sigma}$, contains the covariances between different elements from a vector of random variables. For example, consider a vector of white noise processes:

$$\hat{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

The variance-covariance matrix for this random vector has as its (i, j) element the covariance between the ith and jth elements of the vector. So the $(1, 2)$ element of the variance-covariance matrix for the white noise processes above is $cov(\epsilon_1, \epsilon_2)$. This can be written out fully:

$$\mathbf{\Sigma} = \begin{bmatrix} cov(\epsilon_1, \epsilon_1) & cov(\epsilon_1, \epsilon_2) \\ cov(\epsilon_2, \epsilon_1) & cov(\epsilon_2, \epsilon_2) \end{bmatrix} \equiv \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

The diagonal elements are the variances of the respective elements of the random vector, which are constant over time because of the white noise assumption.[5] Similarly, the off-diagonal elements are the covariances, which are zero because white noise processes are uncorrelated by definition. Variance-covariance matrices of estimators are frequently used, particularly for OLS.

Finally, a convenient way to cut down on notation in time series analysis is to use lag operators. This operator moves a variable back the specified number of periods, i.e.:

$$L^i y_t = y_{t-i} \tag{2}$$

Putting the lag operator, L^i, before a variable y_t will lag that variable by i periods. A negative lag

[5] The covariance of a variable with itself is the same as the variance.

means it will move the variable forward by i periods. As another example, the random walk model from above can be written using a lag operator:

$$y_t - Ly_t = \epsilon_t \quad \equiv \quad A(L)y_t = \epsilon_t \tag{3}$$

where $A(L) = (1 - L)$.

Deterministic and Stochastic Trends

White noise is also a convenient assumption because such a process is stationary. A particular variable y_t is said to be stationary if the sequence of its realizations $\{y_t\}$ yields a constant mean, variance, and autocovariance. Additional realizations of the process do not change its expected value, the spread of the realizations from the mean stays the same, as does the correlation between different realizations at any point in time.

While it may be plausible to restrict error terms to be stationary, most macroeconomic aggregates of interest are not stationary. GDP, consumption, and other important macroeconomic variables grow over time -they have a trend. Trends are differentiated between those which are deterministic and those which are stochastic. Separating an arbitrary stochastic difference equation into three parts is a good way to see this clearly:

$$y_t = \text{trend} + \text{stationary component} + \text{noise} \tag{4}$$

By assumption the stationary and noise components of this process have a constant mean, variance, and autocovariance. Whether the series is stationary overall will depend on the properties of the trend. If no trend exists, as some claim is true of the real oil price, the series is stationary. However, if the series is trending this component can either be deterministic or stochastic. If it is deterministic, the trend can be predicted, and any deviations from the trend will be relatively short-lived. The basic idea is that the impact of noise on the trend dies out over time, so that the realizations of the time series return to the deterministic trend (plus the stationary component). This is called a trend-stationary process. If some long-run growth rate of real GDP exists in the economy, then this growth rate is a trend-stationary process.

A stochastic trend cannot be predicted, and any deviations from the trend may or may not be short-lived. The reason for this is that the impact of the noise on the trend has a permanent effect, and the noise is random by construction, making the trend random as well. This is called a process with a stochastic trend. An example from Kennedy (2008) makes the difference clear. Consider two different

ways of modeling GDP (y_t), where θ is the growth rate of GDP, and ϵ is an error term with mean 1:

$$y_t = y_0 \exp \theta t \epsilon_t \tag{5}$$

$$y_t = y_{t-1} \exp \theta t \epsilon_t \tag{6}$$

Equation (5) predicts that GDP grows exponentially beginning at y_0 from that point forward. The error term does play a role, but only to the extent that it forces GDP to move away from the exponential path. Notice that there is no impact of ϵ_{t-1} on y_t, as y_t depends only upon the exponential trend from y_0. Equation (6) is similar, in that it grows at the same exponential rate, but the trend depends on y_{t-1}, not y_0. The difference is that noise from the previous period, ϵ_{t-1}, does impact y_t. This process may never return to the trend which began at y_0. To make this clearer, take logs of both equations:

$$\ln y_t = \ln y_0 + \theta t + \ln \epsilon_t \tag{7}$$

$$\ln y_t = \ln y_{t-1} + \theta + \ln \epsilon_t \tag{8}$$

Next, to get them both in a similar form, substitute repeatedly for the lagged $\log y$ term on the right-hand side of equation (8). With this substitution, the equations are:

$$\ln y_t = \ln y_0 + \theta t + \ln \epsilon_t \tag{9}$$

$$\ln y_t = \ln y_0 + \theta + \Sigma_{i=1}^{t} \ln \epsilon_t \tag{10}$$

Equation (7) has two stationary components (y_0 and the error), and one trending component (θ_t). The trend is deterministic, in the sense that it grows at the constant rate of θ over time without permanently impacting current GDP. This is because past values of the error term have no impact on y_t. Equation (10) also has two stationary components (y_0 and θ), but the error terms are very different. In this equation, past values of the error term do impact y_t, so that shocks to GDP are permanent. A seminal paper by Nelson and Plosser (1982) showed that most macroeconomic aggregates of interest are not trend-stationary, but appear to have a stochastic trend. The growth rates of these aggregates, however, may be stationary.

Models with stochastic trends, such as equation (10) or the random walk above, are said to have

unit-roots. This comes from the fact that the coefficient on the $\ln y_{t-1}$ term in equation (8) is 1. Another name for a process with a stochastic trend is a difference stationary process. This is because differencing many stochastic processes will lead to a trend stationary process. Take for example the random walk from above:

$$y_t = y_{t-1} + \epsilon_t \tag{11}$$

Subtract y_{t-1} from both sides to give:

$$\Delta y_t = \epsilon_t \tag{12}$$

Because the error term is stationary, so is the first difference ($\Delta y_t = y_t - y_{t-1}$)) of the y_t, i.e. the differenced process is stationary, or difference stationary. In general, a non-stationary series can be made stationary by differencing. If the series needs to be differenced once to be made stationary (first differenced) it is called integrated of order one or I(1), if it needs to be differenced twice it is I(2), and so forth. A stationary series is I(0). A trend-stationary series can be made stationary without differencing by removing the deterministic trend, although finding this trend may not be straightforward. As an example, equation (7) grows by a deterministic trend (θ) over time. If such a trend can be found, removing it from the data is relatively straightforward.

Care must be taken to remove a trend in the appropriate manner. If a deterministic trend is removed from a model with a stochastic trend, the series is still not stationary. To see this consider a random walk with drift:

$$y_t = y_{t-1} + \mu + \epsilon_t \tag{13}$$

The only addition to a random walk here is the constant μ. Through repeated substitution for y, this can be rewritten as:

$$y_t = y_0 + \mu t + \Sigma \epsilon_t \tag{14}$$

Even if the deterministic trend is taken out, μt, previous error terms will impact the current value of y. Without some other transformation, the series remains a stochastic process. This means that the variance of y grows over time, which is problematic for test statistics. Trying to difference a trend-stationary series causes problems as well. In this case, another non-stationary process may be introduced into the series. A third approach to removing a trend is to use a filter designed to separate trend from cycle. Popular filters include the Hodrick-Prescott (HP) filter and various Band-Pass (BP) filters (Enders, 2010).

Unit Root Tests

A simple way to look for a unit root (non-stationarity) in a time series process is to inspect the autocorrelation function. This is often called a correlogram, and is a plot of the autocorrelations of a process over time. That is, it plots $corr(y_t, y_{t-1})$,$corr(y_t, y_{t-2})$, ..., $corr(y_t, y_{t-n})$. If the series is stationary, this should go to zero quickly, meaning that past values of y do not have an important impact on the current values. A slow decay in the correlogram over time is a good indication that the process is not stationary.

This visual inspection can be useful, but interpretation of the correlogram may differ among analysts. It may also be difficult to differentiate between a stationary series, and one that is nearly stationary. For these reasons many formal tests for unit roots have been in time series have been developed. The most common unit root test is the Dickey-Fuller (DF) test, which is outlined here following Enders (2010). Other unit root tests are covered in detail in Enders (2010) as well.

Assume that a particular series y_t is generated from the following first-order process:

$$y_t = a_1 y_{t-1} + \epsilon_t \tag{15}$$

In this equation y is a variable, a a coefficient, and ϵ a white-noise process. If $|a| < 1$, then y is stationary and a_1 can be efficiently estimated by OLS. One can then conduct hypothesis testing using the t-statistic generated from the estimation. The standard hypothesis which is tested asks whether a_1 is significantly different from zero. Recall that a unit root process has $|a| = 1$, so to test for a unit root one might want to ask if a_1 is significantly different from one. In theory, rejection of this hypothesis indicates that the series is either stationary ($|a| < 1$) or explosive ($|a| > 1$).

However, using a null hypothesis of non-stationarity raises some conceptual difficulties. In this case y_t is assumed to be non-stationary (this is the null hypothesis), so must be generated by a non-stationary process such as:

$$y_t = y_0 + \sum_{i=1}^{t} \epsilon_i \tag{16}$$

The final term in this equation shows that the error term has a permanent impact on the current value of y. As explained above, one implication of non-stationarity is that the variance of y_t becomes infinitely large as t increases. Because of this it is inappropriate to use t-tests to perform significance tests on the a coefficient. The DF test is a means of overcoming this issue.

Assuming that y_t is a non-stationary series, the DF test subtracts y_{t-1} from both sides of equation

(15) which gives:

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t \tag{17}$$

This modified equation can be used to test for a unit root in y_t. Because $\gamma = (a_1 - 1)$, the null hypothesis that γ=0 is equivalent to the null hypothesis that a_1=1. Because of the structure of this model the t-statistic estimated from the model does not have a standard distribution, and so a distribution specific to the test must be used for the hypothesis test.

The version of the DF test shown here only allows for one lag. The Augmented Dickey-Fuller (ADF) test extends this to allow for multiple lags in the variable under consideration.

Cointegration

While differencing an I(1) process can make it stationary, this can entail some costs in a multivariate setting. In particular, some non-stationary variables tend not to drift too far apart because there are forces that keep them together. In the simplest (and most common) case when such variables are individually I(1) but a linear combination of them is I(0), they are termed cointegrated. Think of consumption and disposable income, imports and exports, and many other important macroeconomic relationships. Differencing each of these variables individually can throw away important information on this relationship.

The major implication of a cointegrating relationship is that differencing and filtering are not the only ways to eliminate unit roots. If a cointegrating relationship between two variables is found, an error correction framework can be used (error correction models are discussed below). Most tests for cointegration between variables take the form of a unit root test applied to the residuals from estimation of the cointegrating relationship. That is, one of the variables in the cointegrating relationship is regressed on the others. If a cointegrating relationship exists the residuals from this regression should be I(0) (Kennedy, 2008).

The following example from Enders (2010) illustrates some key features of a cointegrating relationship. Begin with a set of economic variables (x_t) in long-run equilibrium, where the β_t are coefficients:

$$\beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_n x_{nt} = 0 \tag{18}$$

Using this definition of an equilibrium, the equilibrium error (e_t) can be defined as:

$$\beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_3 x_{3t} = e_t \tag{19}$$

For the equilibrium to be meaningful, the error process (e_t) must be stationary. If the process is not

stationary, then there is no equilibrium point which can be used as a reference for the relationship between the variables. More formally, the x_t are said to be cointegrated of order d, b, denoted by $x_t \sim CI(d, b)$, if:

- All components of x_t are integrated of order d.
- There exists a vector of the β_t such that the linear combination $\beta x_t = \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_n x_{nt}$ is integrated of order $(d - b)$, where b >0.

The vector of the β_t is called the cointegrating vector. The basic idea is that the x_t are integrated of a certain order, and their linear combination is stationary. So if the x_t are individually I(1), and their linear combination is I(0), then the x_t are said to be cointegrated of order (1,1). This usually applies to pairs of variables, but the definition is general enough to encompass cointegrating relationships between more than two variables. This general definition has several other implications, which are covered in detail in Enders (2010).

Testing for Cointegration

If one suspect that two or more variables are cointegrated, there are three different ways of testing for cointegration. Single-equation tests are outlined in this sub-section, while details on VAR and error-correction tests can be found in Kennedy (2008). The most common single-equation test follows the procedure of Engle and Granger.

Assume that two series (x_t and y_t) are believed to be cointegrated. Enders (2010) recommends implementation of the Engle-Granger methodology in the following manner. Begin by testing for the order of integration of each variable. They must be integrated of the same order to have a cointegrating relationship. This is also true if there are more than two variables being tested. The next step is to estimate the long-run relationship, given that the two variables are both integrated of the same order (with that order greater than 0). In the two-variable case this takes the form:

$$y_t = \beta_0 + \beta_t z_t + e_t \tag{20}$$

The residuals from this equation ($\hat{e}_t$) contain estimates of the deviations from the long-run relationship in each period. The final step is then to use a unit root test to determine the order of integration of the residuals. If these deviations are stationary, this indicates that the two variables are moving around some equilibrium value, whereas a unit root in the residuals indicates no long-run relationship.

This method, while relatively straightforward, runs into problems when dealing with more than two variables, and in deciding whether y or z should be on the right or left-hand-side of equation (20). Other methods outlined in Kennedy (2008) attempt to improve on these weaknesses.

Basic Time Series Models

This section moves from basic time series concepts to discuss specific models in more detail. It begins with univariate autoregressive moving average (ARMA) models. The second section extends these to additional variables by considering vector autoregressive models (VARs). The third section describes a modification of these methods, the error-correction model, which is widely used in macroeconometric modeling.

ARMA Models

Autoregressive moving average (ARMA) models are very important in time series analysis, and form the basis for vector autoregressions. They are a combination of both autoregressive processes and moving average processes. An autoregressive process with p lags [$AR(p)$] is one where a variable (y_t) is dependent on only its own p lags and a disturbance term. For example, an AR(2) process is written:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + c_0 \epsilon_t \tag{21}$$

where a_0, a_1, a_2, and c_0 are unknown coefficients to be estimated and the ϵ_t are Gaussian white noise. A moving average process with q lags [$MA(q)$] is written only in terms of the disturbances, with q lags. For example, an MA(2) is written:

$$y_t = \sum_{i=0}^{2} c_i \epsilon_{t-i} \tag{22}$$

An ARMA process is a combination of autoregressive and moving average processes. An $ARMA(p, q)$ is a combination of an $AR(p)$ and $MA(q)$:

$$y_t = a_i y_{t-1} + ... + a_p y_{t-p} + c_i \epsilon_{t-1} + ... + c_q \epsilon_{t-q} \tag{23}$$

Alternatively, the individual autoregressive or moving average process are special cases of the ARMA process. In this case one can think of an $MA(q)$ as an $ARMA(0, q)$, and an $AR(p)$ as an $ARMA(p, 0)$.

Vector autoregressions often use the fact that stationary $AR(1)$ processes can be represented as an $MA(\infty)$. That is, there is a relationship between autoregressive and moving average processes. This can be shown by starting with an arbitrary $AR(1)$ process at t:

$$y_t = a_i y_{t-1} + c_i \epsilon_t \tag{24}$$

This same process is assumed to hold over all periods, so it can be moved back to get an equation for y_{t-1}:

$$y_{t-1} = a_{i-1}y_{t-2} + c_{i-1}\epsilon_{t-1} \tag{25}$$

This can be used to substitute equation (25) into equation (24):

$$y_t = a_i(a_{i-1}y_{t-2} + c_i\epsilon_t) + c_{i-1}\epsilon_{t-1} \tag{26}$$

The procedure is continually repeated for y_{t-2}, y_{t-3}, ..., y_{t-n}. Notice how each successive substitution adds an additional error term to the equation, moving it closer to a moving average representation. These substitutions also add the product of the coefficients on the lagged variables ($a_i a_{i-1}$ in equation (26) above). In order to write this $AR(1)$ as a moving average process, this product must go to zero as the series goes further and further back into the past. Another way to say this is that as $n \to \infty$, $a_i, a_{i-1}, ..., a_{i-n} \to 0$.

The reason that only a stationary $AR(1)$ process has an equivalent $MA(\infty)$ representation is that the product of coefficients only goes to zero if the coefficients are less than one. This is a necessary condition for the process to be stationary. The interpretation of the coefficient being less than one is that the impact of past y values decreases over time. That is, lagged values of y do not have permanent impacts on the value of y far into the future. Taking the limit of the successive iterations from above gives:

$$y_t = \sum_{i=0}^{\infty} c_i\epsilon_{t-i} \tag{27}$$

The interpretation of the coefficient values from this representation is important when using vector autoregressions. Each c_i summarizes the impact of a one unit movement in ϵ_{t-i} on the current value of y_t. For example, c_0 gives the impact of a one unit movement in the current disturbance term on the current value of y_t. This is often called the instantaneous impact.

This process can also be written using the lag operator:

$$y_t = B(L)\epsilon_t \tag{28}$$

where $B(L) = c_0 + c_1L + c_2L^2 + ...$

Vector Autoregressive Models

In a vector autoregression, the variables of interest (endogenous variables) form a vector. It is assumed that each of these endogenous variables impacts the others, possibly simultaneously. This relationship is summarized in the structural representation of a VAR, which postulates that this vector of endogenous variables can be approximated by a vector autoregression of order p. For the two variable, first-order case this reads:

$$b_{11}y_t = b_{10} - b_{12}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \epsilon_{1t} \tag{29}$$

$$b_{22}z_t = b_{20} - b_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \epsilon_{1t} \tag{30}$$

It is important to fully understand each element of these two equations. The first point to note is that the existence of such a linear relationship is in itself an assumption. One should always ask if this is a reasonable assumption. In this case, the endogenous variables are y and z, say GDP and the oil price. Notice how the current values of each endogenous variable are allowed to impact the other, making this a structural representation. Because there is only one lag of each variable in either equation, the system is first-order. The coefficients summarize the impact of each variable on the other. For example, $-b_{21}$ is the contemporaneous impact of a unit change in y_t on z_t, and $-b_{12}$ is the contemporaneous impact of a unit change in z_t on y_t.

The ϵ_t may be referred to as structural innovations or structural shocks or structural disturbances, and are also termed residuals as well. When using a VAR to study the impact of one variable on another (sometimes broadly referred to in the literature as policy analysis) these are the primary objects of interest. They are assumed to be a white noise processes. Due to these assumptions, the shocks represent unexpected movements in either y_t (ϵ_{1t}) or z_t (ϵ_{2t}). Technically, they do not have to be specifically related to a specific variable, although they are often interpreted in this way. Rather, they are the causes of unexpected movements in the value of that variable which are unpredictable and uncorrelated with other endogenous variables or innovations.

VARs are generally estimated using OLS (discussed in the next section), which requires transforming the structural equations to reduced form. This is because estimation by OLS requires the right-hand side variables be uncorrelated with the error term. A quick look at equations (29) and (30) shows this is not the case if current values of either endogenous variable remain on the right-hand side. For the two variable, first-order VAR above the reduced form representation is given by:

$$\underbrace{\begin{bmatrix} 1 & b_{12} \\ b_{21} & 1 \end{bmatrix}}_{\mathbf{B}} \underbrace{\begin{bmatrix} y_t \\ z_t \end{bmatrix}}_{\hat{x}_t} = \underbrace{\begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix}}_{\hat{\Gamma}_0} + \underbrace{\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix}}_{\mathbf{\Gamma}_1} \underbrace{\begin{bmatrix} y_{t-1} \\ z_{t-1} \end{bmatrix}}_{\hat{x}_{t-1}} + \underbrace{\begin{bmatrix} \epsilon_{yt} \\ \epsilon_{zt} \end{bmatrix}}_{\hat{\epsilon}_t} \tag{31}$$

Notice here that both b_{11} and b_{22} have been normalized to equal one, which is standard. More succinctly:

$$\mathbf{B}\hat{x}_t = \hat{\Gamma}_0 + \mathbf{\Gamma}_1\hat{x}_{t-1} + \hat{\epsilon}_t \tag{32}$$

Pre-multiplication by $\mathbf{B}^{-1}$ yields the VAR in reduced (or standard) form:

$$\hat{x}_t = \mathbf{A}_0 + \mathbf{A}_1\hat{x}_{t-1} + \hat{e}_t \tag{33}$$

Where $\mathbf{A}_0 = \mathbf{B}^{-1}\hat{\Gamma}_0$, $\mathbf{A}_1 = \mathbf{B}^{-1}\mathbf{\Gamma}_1$, and $\hat{e}_t = \mathbf{B}^{-1}\hat{\epsilon}_t$. Take a close look at this general representation. The main difference from the structural form is that current values of either endogenous variable are no longer on the right-hand side of the equation. Each matrix has also been transformed by pre-multiplication, including the structural innovations. This is the key point: the reduced form no longer directly represents the structural shocks. Rather, the reduced form is based on a transformation of the structural shocks, namely $\hat{e}_t$. This becomes clearer if written out fully:

$$y_t = a_{10} + a_{11}y_{t-1} + a_{12}z_{t-1} + e_{1t} \tag{34}$$

$$z_t = a_{20} + a_{21}y_{t-1} + a_{22}z_{t-1} + e_{2t} \tag{35}$$

The items of interest for policy analysis, the structural shocks, are no longer directly represented here. They have been replaced by error terms (e_{1t} and e_{2t}). Also notice again how this system has each endogenous variable dependent only on lags of itself and lags of the other endogenous variable. Because of this, the two variable first-order system can be estimated by OLS equation-by-equation to yield coefficient estimates and associated error values. The validity of the estimates, however, depends upon the assumptions underlying OLS holding. Macroeconometric models can be written as VARs, and are also generally estimated equation-by-equation for the same reason as VARs.

Error Correction Models

An error correction model (ECM) is designed to incorporate long-run information available from cointegrated variables in addition to any relationship implied by the sample data. This type of model can be illustrated using an example from Kennedy (2008) for the two variable case, where the B are coefficients and e_t the error term:

$$y_t = B_0 + B_1x_t + B_2x_{t-1} + B_3y_{t-1} + e_t \tag{36}$$

This standard equation regresses y_t on x and lags of both x and y. Now suppose that both x and y are individually non-stationary and economic theory suggests that in the long-run they are cointegrated, with relationship $y = \psi + \theta x$. Ideally, one would like to use the information contained in the cointegrating relationship in describing the relationship, but short and long-term, between these two variables. Using an ECM framework is one way this can be done. Equation (36) can be manipulated to yield an ECM (see Kennedy (2008) for the derivation):

$$\Delta y_t = B_1 \Delta x_t + (B_3 - 1)(y_{t-1} - \psi - \theta x_{t-1}) + e_t \tag{37}$$

Now the first-differences of y move proportionally to the first-differences of x, as well as the long-run relationship. It is in this sense that an error-correction model is able to incorporate long-run information from cointegrated variables. The second term on the right-hand side is called the error correction term, and adjusts for any deviations from the long-run relationship. For example, if y_{t-1} strayed relatively far from x_{t-1} then $y_{t-1} > \psi - \theta x_{t-1}$. But because B_3-1 is negative ($B_3 < 1$ due to stationarity), Δy_t is reduced to bring it in-line with the long-run relationship.

The strength of the ECM model is that it uses both differenced variables as well as levels values. In theory this should give it an advantage in generating coefficient estimates. The ECM approach also allows the incorporation of economic theory into the equation specification, as theory can be used to generate the error correction term. This ECM can be extended to a vector error correction model (VECM) in the same manner univariate regressions are extended to VARs. In the presence of cointegrated variables, a VECM can be used to generate forecasts rather than a VAR with transformed data.

For any set of I(1) variables, the Granger Representation Theorem says that error correction and cointegration are equivalent representations. An example from Enders (2010) helps to illustrate this point. Consider an error-correction model with two equations and two endogenous variables:

$$\Delta r_{St} = \alpha_S (r_{Lt-1} - \beta r_{St-1}) + \epsilon_{St} \quad \alpha_S > 0 \tag{38}$$

$$\Delta r_{Lt} = \alpha_L (r_{Lt-1} - \beta r_{St-1}) + \epsilon_{Lt} \quad \alpha_L > 0 \tag{39}$$

The endogenous variables in this model are the short-term interest rate (r_{St}) and the long-term interest rate (r_{Lt}). The ϵ_{St} and ϵ_{Lt} are white-noise disturbances and the α and β are coefficients. Each of the equations is specified in error-correction form, with the I(0) difference on the left-hand side and the error-correction term on the right-hand side. Because the left-hand side is stationary by assumption in either equation, this must be true of the right-hand side has well. As the only term on the right-hand side is the error-correction one, this implies that the linear combination of these variables is stationary,

or that they are cointegrated. Enders (2010) provides additional detail on more complex examples.

Estimation

This section moves to the estimation of models with both stationary and non-stationary variables. Multi-variate ordinary least squares (OLS) is reviewed first, including the estimation of ECMs by OLS. This is followed by an overview of two-stage least squares, which is sometimes used in estimating macroeconometric models.

Ordinary Least Squares

Recall that in the time series context, a general OLS equation with n observations and two independent variables is written as:

$$y_t = B_0 + B_1 x_t + B_2 z_t + u_t \quad t = 1, 2, ..., n \tag{40}$$

Here, B_0 is a constant, B_1 is a coefficient which gives the impact of the independent variable x_t on the dependent variable y_t, B_2 is also a coefficient with a similar interpretation for z_t, and u_t is the error term.

This equation can also be written in vector form:

$$y_t = \hat{J}_t \hat{B} + u_t \quad t = 1, 2, ..., n \tag{41}$$

where $\hat{J}_t = (1, x_t, z_t)$ is a 1x3 vector of independent variables, and $\hat{B} = (B_0, B_1, B_2)'$ is a 3x1 vector of coefficients. The notation can be consolidated further in matrix form if the n observations are included:

$$\hat{y} = \mathbf{J}\hat{B} + \hat{u} \tag{42}$$

Now $\hat{y}$ is the nx1 vector of observations of the dependent variable, $\hat{B}$ is still the 3x1 vector of coefficients, and $\hat{u}$ is the nx1 vector of unobserved errors. The nx3 matrix **J** contains the observations of independent variables, where each row corresponds to one observation and the columns are the independent variables. The method of ordinary least squares provides estimates of the constant and coefficient values by minimizing the sum of squared errors from the system above.

There are five basic assumptions made when using OLS, which must also hold when estimating VARs or macroeconometric models. First, it is assumed that the dependent variable can be calculated as a linear function of the independent variables plus a disturbance term. For VARs, this assumption is slightly modified in that the endogenous variables are assumed to be linear functions of the other

endogenous variables. Second, the expected value of the disturbance term is zero. Third, the disturbance terms have constant variance and are uncorrelated with each other.

Fourth, the observations of the independent variables can be considered fixed in repeat samples. This assumption is often weakened so that it is required only that the independent variables (or variables on the right-hand side of the regression) are uncorrelated with the error term. Finally, it is assumed there are more observations than independent variables and there are no exact linear relationships between independent variables.

If these assumptions are believed to hold, estimating parameters by OLS requires minimizing the sum of squared residuals. This amounts to writing equation (40) in terms of estimates:

$$y_t = B_0^{OLS} + B_1^{OLS} x_t + B_2^{OLS} z_t + e_t \quad t = 1, 2, ..., n \tag{43}$$

where the B^{OLS} indicate estimates, and e_t is the error of the estimate each time period, sometimes called the residual. Rearrange this equation and square to get the sum of squared residuals:

$$SSE = \sum_{t=1}^{n} (y_t - B_0^{OLS} + B_1^{OLS} x_t + B_2^{OLS} z_t)^2 \tag{44}$$

The equation is squared to remove the influence of negative values. The next step is to choose B_0^{OLS}, B_1^{OLS}, and B_2^{OLS} to minimize this term. This leads to three equations (called first-order conditions) for each time period, which can be solved simultaneously to give the desired estimates. As an example, the coefficient B_1^{OLS} has the following formula after solving the system of equations:

$$B_1^{OLS} = \frac{\sum_{t=1}^{n} (x_t - \bar{x}) y_t}{\sum_{t=1}^{n} (x_t - \bar{x})} \tag{45}$$

The other two coefficients will have a similar formula. This procedure can be generalized with matrices in a similar manner. The formula resulting from minimizing the sum of squared errors in this case is:

$$\hat{B} = (\mathbf{J}' \mathbf{J})^{-1} \mathbf{J}' \hat{y} \tag{46}$$

Each of the three coefficients is summarized in the vector $\hat{B}$.

Problems arise in hypothesis testing with OLS estimation if the regressors are non-stationary. Specifically, if one of the regressors in an equation has a stochastic trend then the t-statistic associated with its OLS estimates can have non-standard distributions (Stock and Watson, 2007). The implication is that the estimated standard errors may be poor approximations to the true standard errors, which means that hypothesis testing may not be valid.

To overcome this limitation a common strategy is to make regressors stationary through differencing, de-trending, or filtering. Another option is to compare the estimated t-statistics with a bootstrap procedure, and to use them if the asymptotic distribution seems appropriate. A final approach, if the relevant variables are cointegrated, is to estimate the equation in ECM form.

Specifically, if the estimating equation contains non-stationary variables, but either all the variables or a subset of them are cointegrated, then an error-correction model can also be estimated. As an example, consider again equation (38) from above:

$$\Delta y_t = B_1 \Delta x_t + (B_3 - 1)(y_{t-1} - \psi - \theta x_{t-1}) + e_t$$

OLS can be used to estimate the coefficients in this equation by two asymptotically equivalent methods. The Engle-Granger two-step procedure proceeds by first estimating the long-run relationship:

$$y_t = \psi + \theta x_t + \epsilon_t \tag{47}$$

The residuals from this equation ($r_t = y_t - \psi - \theta x_t$) can then be put into equation (38) and then the entire equation can be estimated:

$$\Delta y_t = B_1 \Delta x_t + (B_3 - 1) r_t + e_t \tag{48}$$

The second method is to estimate the coefficients in one-step by slightly rewriting equation (38):

$$\Delta y_t = B_1 \Delta x_t + (B_3 - 1)(y_{t-1} - \psi - \theta x_{t-1}) + e_t = -(B_3 - 1)\psi + B_1 \Delta x_t + (B_3 - 1) y_{t-1} - (B_3 - 1)\theta x_{t-1} + e_t \tag{49}$$

Two-Stage Least Squares

Because macroeconometric models are systems of simultaneous equations with many endogenous variables, it is likely the case that the right-hand side variables are correlated with the error terms, resulting in a violation of OLS assumption four from above. A popular method for overcoming this violation is to use two-stage least squares estimation instead of OLS.

As the name implies, the estimation occurs in two stages. In the first stage each endogenous variable acting as a regressor in the equation is estimated on all of the exogenous variables in the system of simultaneous equations. This yields estimated values for each endogenous variable in the particular equation. In the second stage these estimated values (instead of the actuals) and the exogenous variables specific to the particular equation are used as regressors to estimate coefficient values.

The idea behind two-stage least squares is to create the best possible instrumental variable for those endogenous variables on the right-hand side of the equation which are not independent of the error term.[6] In two-stage least squares this is done by combining all of the exogenous variables (in stage 1). The estimate for each endogenous variable generated in stage 1 is then highly correlated with the regressor for which it is acting as an instrumental variable, but independent of the error terms in the equation being estimated.

Estimates for non-stationary regressors in two-state least squares are dealt with in the same manner as with OLS.

The Structure of Macroeconometric Models

This section surveys the common structure of macroeconometric models. Although the number, type, and theory behind equations differs between macroeconometric models there is a general specification structure. Macroeconometric models also have similarities in how they are estimated, usually either by OLS or two-stage least squares. As with estimation, solution of these models can be fairly standard through use of the Gauss-Siedel algorithm, although this can differ for models with forward looking variables. This section also outlines methods used to test model equations and results, and outlines a few additional factors that can be important in using and interpreting macroeconometric models.

Specification

The specification of macroeconometric models often is often idiosyncratic to the modeler. It can be based upon theory, experience and judgement, various statistical tests, or a combination of each. This section follows Chapter 1 of Fair (2012) and highlights some key features of macroeconometric model specification which are common across models.

Basic Specification

Macroeconometric models are systems of stochastic (or behavioral) equations and identities. Stochastic equations are usually based at least in part on theory and are estimated using historical data, whereas identities are equations that hold by definition. These equations are comprised of both endogenous and exogenous variables. The values of endogenous variables are determined in the model, they are the solution to the model equations. Exogenous variables come from outside of the model structure.

Although all macroeconometric models begin with this basic framework, they will differ in the particular theory which is used in each stochastic equation, whether the equations are written in levels

[6] An instrumental variable is one which is highly correlated with the variable it is intended to replace, but which does not impact the dependent variable.

or in error correction form, the split between endogenous and exogenous variables, and in their size and scope.

An example helps to make this clear. In the equations below (where t is the time period) there are three endogenous variables, consumption (C_t), investment (I_t), and income (Y_t). There are also two exogenous variables, government spending (G_t) and the interest rate (r_t). The three equations are:

$$C_t = a_1 + a_2 Y_t + e_t \tag{50}$$

$$I_t = b_1 + b_2 r_t + u_t \tag{51}$$

$$Y_t = C_t + I_t + G_t \tag{52}$$

The a_i and b_j are coefficients which summarize the relationships between variables, these are what we use the data to estimate. And the e_t and u_t are error terms; they encompass all of the other variables which help to explain the endogenous variables but are unaccounted for in the equations.

Equations (50) and (51) are examples of stochastic equations, and their specification (i.e. the right-hand side) can be based on theory, judgement, statistical fit, or other factors important to the modeler. Equation (50) is the consumption function, and states that current consumption depends on current income. The investment function, equation (51), says that investment depends on the interest rate. Equation (52) is an example of an identity, in this case the national income identity. Notice that it does not have an error term because this equation must hold by definition.

These simple equations are unlikely to explain or predict macroeconomic phenomenon very well, and so require modification. One issue is that each of the stochastic equations above contains only current values of the variables on the right-hand side. In most cases stochastic equations will have lags of variables on the right-hand side as well. For example, equation (50) can also be written:

$$C_t = a_1 + a_2 Y_t + a_3 C_{t-1} + a_4 Y_{t-1} + e_t \tag{53}$$

In this case consumption depends on the current value of income, but also the value of consumption and income last period. These backward looking expectations are the most common specification in macroeconometric models, but forward looking expectations are also used. For example, equation (50) with forward looking expectations can be written:

$$C_t = a_1 + a_2 Y_t + a_3 E_t Y_{t+1} + e_t \tag{54}$$

In this case E_t is short-hand for the conditional expectation of Y_{t+1} given the information at t, or $E_t y_{t+j} = E(y_{t+j}|y_t, y_{t-1}, ...)$. The specification used here indicates that current consumption depends on current income and expected future income.

One might also want to add more than one lag, additional endogenous or exogenous variables, or multiple lags of one variable and only one lag of another. For each stochastic equation in the macroeconometric model such as equation (53), there are many different specifications. This is also true for the overall specification of the model itself, although there is a general structure which tends to be implemented.

General Structure

The overall structure of macroeconometric models often follows a top-down approach. Major aggregates of interest are estimated using specific stochastic equations, and the remainder of variables are computed using identities or simple relations with the major aggregates. The definition of "major aggregates" differs between each model, and often depends on the interests of the modeler and purpose of the modeling exercise.

As an example, consider again equation (51). Because, we are often interested in the details of investment (and not just the aggregate value), this equation could be replaced by three different equations. The model could use stochastic equations for residential ($I_{r,t}$) and non-residential investment ($I_{nr,t}$):

$$I_{r,t} = b_{r,1} + b_{r,2} r_{m,t} + u_{r,t} \tag{55}$$

$$I_{nr,t} = b_{nr,1} + b_{nr,2} r_{c,t} + u_{nr,t} \tag{56}$$

In this case residential investment is specified as a function of the interest rate on mortgages ($r_{m,t}$) whereas non-residential investment depends on the yield on commercial bonds ($r_{c,t}$). These simple equations can be altered as above by adding lags and different exogenous and endogenous variables. Given these two stochastic equations, the third equation determines total investment through the identity:

$$I_t = I_{r,t} + I_{nr,t} \tag{57}$$

The same type of procedure can be used to disaggregate consumption, government spending, exports, and imports. In fact, many macroeconometric models begin with the national income identity ($Y_t = C_t + I_t + G_t + NX_t$) and disaggregate each component in the same manner as above. The art

of building such models is in determining at which level the stochastic equations should appear, how much detail is necessary within each macroeconomic aggregate, and the specific variables to use within each stochastic equation. It is easy to see why macroeconometric models grow quickly in size and complexity.

This general structure also highlights major strengths and weaknesses of macroeconometric models. Because they can be expanded relatively easily, such models are very useful for obtaining detailed (and consistent) forecasts/projections of various variables of interest. And these variables of interest often line up exactly with the national accounts.

This level of detail, however, can make such models difficult to interpret and use. It can be unclear what is driving different forecasts of model results. The top-down approach also is unable to account for interactions between different sectors, as there are no explicit links specifying the input of one sector might be the output of another. Rather, changes to each variable are determined in the stochastic equations at a relatively high-level, and make their way into more detail from this higher level.

Estimation

Once the model has been specified, the next step is to incorporate historical data and estimate the relevant coefficients. Most macroeconometric models are estimated equation-by-equation using OLS. However, the structure of the model and underlying data may need to be changed in order to make this a plausible estimation strategy. This section outlines some considerations necessary for estimation.

Simultaneity

Macroeconometric models are systems of simultaneous equations. One problem that often arises in estimating such systems is simultaneous causality between the variables on the left and right-hand sides of each equation. For example, in the simple system above [equations (50)-(52)] does income determine consumption or the reverse? This so-called simultaneity or simultaneous causality can lead to correlation between the regressor and the error term in each stochastic equation.

To see why this is the case consider again equation (50) from above. Suppose that for some unexplained reason the error term (e_t) rises (perhaps some omitted factor such as wealth is now higher). This directly raises consumption, but because Y_t also depends on consumption [from equation (52)], this changes Y_t as well, which further changes C_t, and so on. The end result is that consumption is correlated with the error term, which violates OLS assumptions, and means that the OLS estimates of this equation are biased and inconsistent.

One way to circumvent this issue is to use two-stage least squares estimation. However, modelers often ignore simultaneity issues, arguing that the bias is not large enough to seriously impair results.

Stationarity

Another issue which affects estimation is that many macroeconomic variables are not stationary. Using OLS for estimation with such trending variables leads to biased test statistics which may invalidate standard hypothesis testing. There are several ways to overcome this issue.

A simple approach is to filter or difference the data before estimation. Filtering is rarely done because it risks throwing away valuable information due to the specific procedure used. Differencing is uncommon for the same reason, if done only for estimation. However, differencing variables for use with an error-correction model is another way around stationarity issues. This requires the variables in each equation to be cointegrated. Because of the size of macroeconometric variables, error-correction type equations are often used with cointegrated variables, and levels specifications with the others. The stationarity concerns remain when estimating these levels equations.

A final option is to ignore the stationarity issues, or to add a deterministic trend to the relevant equations. Neither of these approaches validates test statistics estimated on trending variables, but the use of such statistics can be tested through a bootstrapping procedure for each equation [see Fair (2004)].

Seasonality

Many macroeconomic variables are seasonally adjusted to account for predictable and expected swings throughout the year. If a macroeconometric model is based on quarterly data, the equations can be adjusted to account for this seasonality or the data is seasonally adjusted before estimation.

Solution

Once the coefficient values have been estimated, the model can be solved to yield values for the endogenous variables in each time period consistent with the coefficients. The Gauss-Siedel algorithm is commonly used to solve macroeconometric models. The basic procedure can be generalized into four steps, following Fair (2012):

1. Guess a set of values for the endogenous variables.
2. Using this set of values for the right-hand side variables, solve for the left-hand side variables.
3. Use these left-hand side values as the new right-hand side values, and solve for the left-hand side variables again.
4. Repeat this procedure until the differences between the new left-hand side variables and the old right-hand side variables is within some degree of accuracy.

One implication of this algorithm is that the order in which equations are solved matters for the speed and reliability of model solutions. Because of its iterative nature, the algorithm solves period-by-period

(usually a quarter or year) in an iterative fashion. If the equations contain forward-looking endogenous variables, as with rational expectations models, a more complicated solution algorithm is required. Details of such an algorithm can be found in Taylor (1993). There are also several ways in which this algorithm can be modified to speed convergence, see Fair (2012) for additional details.

Testing the Equations

Macroeconometric models are commonly tested during and after specification, estimation, and solution. In general, single equation tests are performed throughout the model building and estimation stages, while full model tests are performed after the model has been solved.

Single Equation Tests

A common type of single equation test is to add a variable or a set of variables to a particular equation and test for significance. This can take the form of adding a lag or multiple lags (or leads) of endogenous and exogenous variables and testing for statistical significance (t-test), testing for joint significance (F-test), or both. A deterministic time trend can also be added and checked for statistical significance, as can an intercept.

Other single equation tests check for different violations of estimation assumptions. One common test is to look for serial correlation of the error term in individual equations. Heteroscedasticity is another violation of such assumptions which is often tested, as is the normality of errors in each equation. If detected, any of these errors can be fixed relatively easily by adding lags or additional regressors, or by using corrected standard errors.

Some modelers also check individual equations for parameter stability, or structural breaks. There are different variants of such tests which can be employed [see Enders (2010)]. However, once parameter instability is found it is unclear how such a problem should be addressed.

Tests of the Full Model

The simplest and most common way to test models is to see how well their forecasts fit the data. The evaluation of forecasts can be either in-sample or out-of-sample. In-sample evaluation consists of using the first t observations of a sample to compute the respective coefficients. The modeler can then calculate the error for the remainder of observations in the sample, say $x + 50$. One method is to use the coefficients from the x observations to calculate all of the errors. The other is to use the coefficients for the x observations to calculate the forecast error with respect to $x + 1$, then re-estimate the coefficients with the $x + 1$ observations, and calculate the forecast error with respect to observation $x + 2$. This procedure can be done for the remainder of observations in the sample. Out-of-sample forecast errors are calculated with observations which are not currently in the sample, and usually become available only over time.

Whichever method is used, the forecaster must pick a way to summarize the forecast errors. There are multiple methods available for this task, and the choice of the best one will depend on the purpose of the forecast. For example, the accuracy criterion used for forecasting turning points in the business cycle will generally be different than one used for forecasting GDP growth at various horizons.

A popular choice to summarize forecast accuracy is mean absolute deviation (MAD). The is the average of the absolute values of the forecast errors. This is sometimes also called mean forecast error (MAE), and is appropriate when the cost of forecast errors is proportional to the absolute size of the forecast error. It is sensitive to scaling. Root mean square error (RMSE) is also sensitive to scaling, and is the square root of the average of the squared values of the forecast errors. This measure weights large forecast errors more heavily than small ones.

Mean absolute percentage error (MAPE) is not sensitive to scaling, and is the average of the absolute values of the percentage errors. It is appropriate when the cost of the forecast error is more closely related to the percentage error than to the numerical size of the error. Another measure is the correlations of forecasts with actual values. The percentage of turning points criteria is a $0/1$ measure which summarizes if turning points were forecast correctly. Each of these has its variations, and there are other methods as well which may be used for specialty forecasts.

The individual equations can then be changed depending on how the model fits observed data with respect to each variable.

Forecasting and Interpreting Results

Once models have been built and tested, they are commonly used for forecasting and policy analysis. This section reviews some factors to consider in each of these contexts.

Add-Factors

Add-factors are additional factors used in individual equations to assist in forecasting. Specifically, they are residuals in equations which can be modified to account for different variables which may not appear in the equation. To illustrate, consider again equation (51) from above:

$$I_t = b_1 + b_2 r_t + u_t \tag{58}$$

Suppose this equation is estimated over the period 1950-2011, and this yields values for the coefficients (b_1, b_2), and the residuals (u_t). To forecast with this model h periods ahead we would use:

$$I_{t+h|t} = b_1 + b_2 r_{t+h|t} \tag{59}$$

$I_{t+h|t}$ is the forecast of investment h periods ahead made at time t, and $r_{t+h|t}$ is the forecast of the

(exogenous) interest rate variable h periods ahead made at t. Notice there is no error term in this equation, and the values of r_t, b_1, and b_2 are known. Add-factors give the modeler some flexibility by adding an additional variable to this equation in each period (af_t):

$$I_{t+h|t} = b_1 + b_2 r_{t+h|t} + af_{t+h|t} \tag{60}$$

In this example af_t can take on any value specified by the modeler in each future period. A good way to think about this is as an exogenous variable which encompasses many of the relevant (endogenous and exogenous) variables which are omitted from equation (59).

How to choose the values for the add-factors in each future period? A common method is to use historical errors as a guide (u_t). One could use the annual average error, or the average error over specific time periods, or even grow the add-factors in the future at the same rate that the errors grew in the past.

Expectations and the Lucas Critique

Policy shifts can change consumer and firm expectations, and possibly behavior. The difficulty in dealing with these expectational (or behavioral) changes is that past responses may not be a reliable guide to future responses. In macroeconometric models, various coefficients summarize the responses of variables to one another, and the coefficients are estimated based on historical data. The standard argument is that a policy shift invalidates these coefficient values because it inherently changes the relationships between the variables, through either consumer or firm expectational or behavioral changes.

This is called the Lucas critique. The weak-form of the Lucas critique states that consumer and firm expectations change with policy shifts, and the strong form says that consumer and firm behavior changes due to these same policy shifts. There is little disagreement about the theoretical validity of the Lucas critique. However, many question the empirical relevance, especially for forecasting.

The take-away is to be careful in interpreting results from macroeconometric models when simulating large (or very different) policy actions. For example, it is unclear that macroeconometric models are able to provide a good guide to the impacts of the Federal Reserve's recent bouts of quantitative easing, or of the macroeconomic impacts of the recent federal fiscal stimulus. This is only because such policy actions have been relatively rare historically, and so the coefficient estimates in macroeconometric models do not encapsulate the behavioral responses to such policy changes.

The Theory of Macroeconometric Models: Short-Run

Roughly speaking, the short-run structure of most macroeconometric models is based on Keynesian ($IS-LM$) theory so that variable outcomes are demand determined. Supply dominates in the

long-run, which is based on the neo-classical growth model of Solow and Swan (see the next section). This section outlines the short-run theory and through the use of an example model with equations. The illustrative model used has multiple countries, assumes that there is a global benchmark interest rate, and that all countries take their terms of trade as beyond their individual control.

The $IS - LM$ Model and Aggregate Demand

This section begins by outlining the theory represented by the short-run structure of many macroeconometric models. Roughly, this is based on the Keynesian $IS - LM$ framework which is used to determine aggregate demand.

The IS Curve

The IS curve summarizes the relationship between the real interest rate and the level of income that arises in the market for goods and services. It can be derived using the Keynesian Cross, which is depicted in Figure 1. The Keynesian Cross shows graphically when actual and planned expenditure coincide. By definition actual expenditures are the same as income in an economy, and an equilibrium exists when these actual expenditures are the same as the planned expenditures of households, governments, and firms.

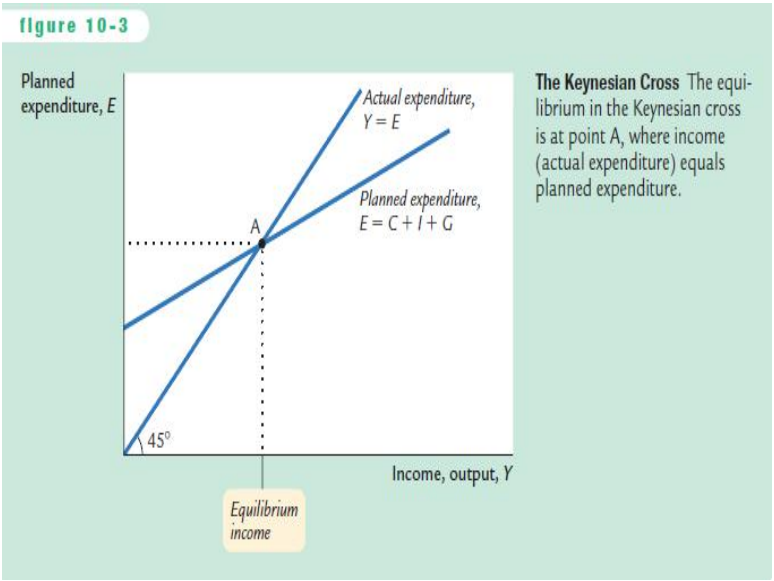


Figure 1: The Keynesian Cross, from Mankiw (2002)

In Figure 1 actual expenditures are a 45 degree line. This is because income equals actual expenditures by definition. Planned expenditures, however, depend on consumption, investment, and government spending (in a closed economy). Many macroeconometric models take government spending as exogenously determined, which means the planned expenditure schedule will shift when this

expenditure is changed. Both the slope and shape of the planned expenditure curve will depend on how consumption and investment change with income. The planned expenditure schedule shown in Figure 1 is linear for simplicity. In actual models, consumption can depend on variables such as income, employment, wealth, and the real interest rate. For example:

$$\Delta c_t = a_1 \Delta y_t + a_2 \Delta u_t - a_3[c_{t-1} - a_4 y_{t-1} - (1 - a_4) W_{t-1} + a_5 R_{t-1}] \tag{61}$$

This equation is in error correction form, where the coefficients (a) are estimated using OLS, lowercase indicates variables are in logs, and the Δ are first differences. The equation says that in any period the percent change in consumption (the log difference) depends directly on the percent change in national income (y) and the percent change in unemployment (u). In addition, the current change in consumption also depends on how the level of consumption deviates from its long run path. The error-correction term in brackets shows that in the long-run consumption depends on the level of national income, national wealth (W), and the real interest rate (R). The benefit of using this type of equation to determine consumption is that it "corrects" the change in consumption when it deviates from the long-run path in any given period.

In this example, assume that each of the variables in the consumption equation are endogenously determined in the model, so the change in consumption must be determined simultaneously with the remainder of model variables and equations. Additionally, there are no forward-looking variables in the consumption equation, so that consumption changes are only backward looking. Another implication of this is that consumption cannot help to derive the relation between the current real interest rate and income in the goods market, which is the IS curve. This must be done through investment.

Many macroeconometric models represent investment generally through private business investment, private housing investment, and government investment. Government investment is often exogenous to the models. Private housing investment is determined in the same way as consumption. Again, because the real interest rate in this equation is lagged, private housing investment cannot contribute to the derivation of the IS curve. Rather, the derivation works through private business investment, which also has an error-correction form (the b are estimated coefficients, lowercase means logs, and Δ the first-difference):

$$\Delta i_t = b_1 q_t + b_2 \Delta y_t - b_3[i_{t-1} - k_{t-1}] \tag{62}$$

Private business investment (i) is directly impacted by changes in national income (y) and the ratio of the marginal return on an investment to its replacement cost (q). This formulation is based on the q-theory of investment, which posits that firms will invest when the benefit of an investment (including taxes and installation costs) exceeds its cost, or $q > 1$. The opposite is true when q is below one. In the long-run this equation forces investment to be proportional to the capital stock (k). Not

shown are accelerator effects on i, which make investment increase or decrease at faster rates based on the level of national income. This helps the model to capture some of the volatility observed in investment over the business cycle.

To derive the IS curve, consider a rise in the current real interest rate. According to the equations, this will not alter consumption, government spending, government investment, or private housing investment. This does, however, reduce planned private investment because the value of q is reduced with a higher real interest rate. The replacement cost of the capital stock rises with a higher real interest rate (higher borrowing costs), but the benefit from that unit does not change. The result is a lower value of q, and an inverse relation between the real interest rate and investment. The fact that the higher real interest rate reduces planned investment shifts down the planned expenditure schedule, which gives a lower level of equilibrium national income.

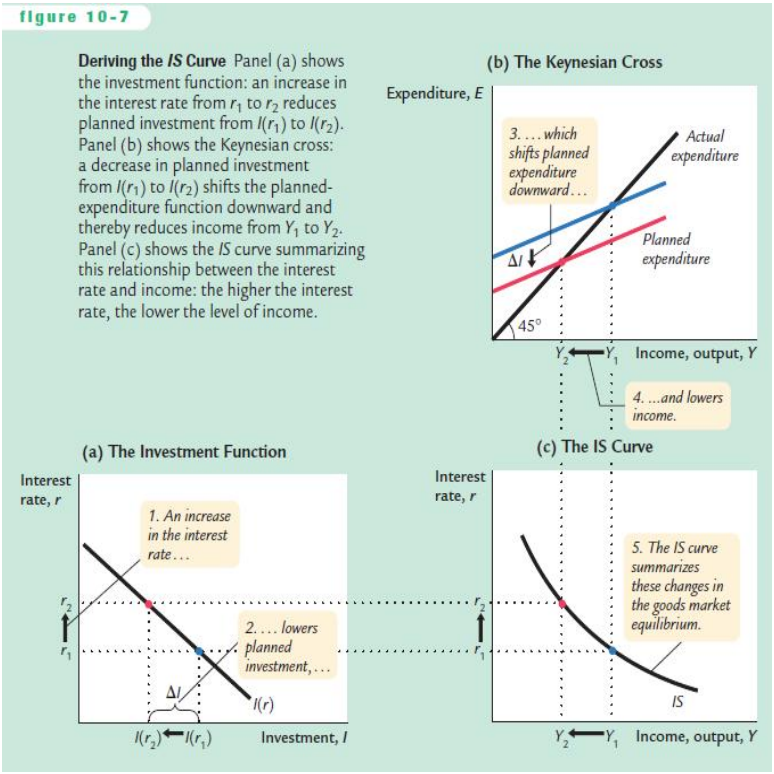


Figure 2: Derivation of the IS Curve Using the Keynesian Cross, from Mankiw (2002)

Figure 2 shows this procedure graphically, and how it results in a relationship between the real interest rate and output in the goods market, or the IS curve. The IS curve shows that a higher real interest rate leads to less income, all else equal, as planned expenditures are lower due to the higher real interest rate. Shifts in the IS curve are also driven by exogenous changes. The curve shifts up with an increase in government spending, a rise in government investment, higher housing investment, greater wealth (which raises consumption), and a decline in taxes (on consumer or firm income). The taxes

are not shown, but can easily be incorporated into the consumption and investment equations.

The LM Curve

The LM curve plots the relationship between the real interest rate and the level of income that arises in the market for money balances. As with the IS curve, the LM curve is derived from an underlying relationship, in this case based on the supply and demand for real money balances. A crucial point here is that these are real money balances ($\frac{M}{P}$), where M is the supply or demand for money and P is the general price level. This means that the LM curve is based on a given, fixed, price level.

The supply of real money balances is assumed to be controlled by a monetary authority. The monetary authority chooses money supply in many models based on a Taylor rule. That is, the monetary authority is assumed to have some target for both inflation and output, and responds to deviations from these targets by changing the supply of real money balances to alter the real interest rate. To build up to this rule, the simplest step is to assume a vertical money supply function. This says that the monetary authority does not consider the real interest rate in choosing the money supply. While not exactly right, it is workable as an approximation.

The other side of this is money demand. Taking income as given, this is assumed to be inversely related to the real interest rate. As the real interest rate is the opportunity cost of holding money for a given income level, a higher real interest rate leads to lower cash holdings. This money market is shown in panel (a) of Figure 3.

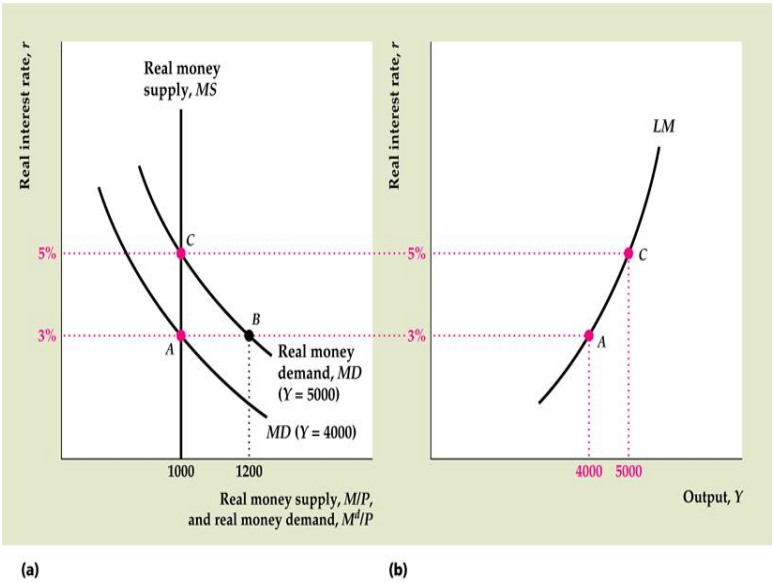


Figure 3: Derivation of the LM Curve Using the Money Market, from Abel et al. (2007)

The LM curve is derived by changing the level of income and finding the new equilibrium real interest rate. A higher level of income increases the demand for real money balances, which shifts out the

money demand curve as shown in panel (a) of Figure 3. In equilibrium this leads to a higher real interest rate, which gives an upward sloping LM curve shown in panel (b) of Figure 3. The LM curve shifts with anything that changes money supply or demand. Examples include a change in the price level, a variation in wealth, changes in the risk of assets other than money, or even alternative payment technologies.

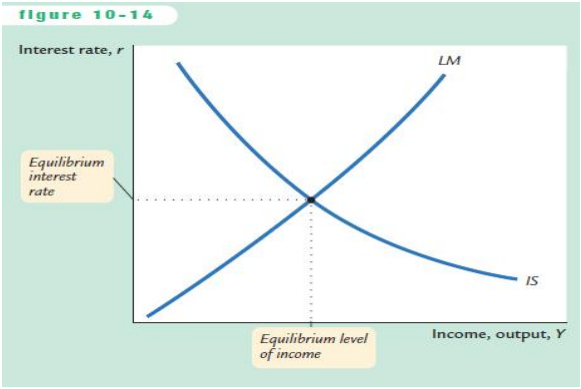


Figure 4: The $IS - LM$ Model, from Mankiw (2002)

The standard $IS - LM$ relationship is depicted in Figure 4. This shows that the equilibrium real interest rate and output are found when both the goods and money markets are in equilibrium. The result will be slightly different in models where monetary policy follows a Taylor rule. This means that the real interest rate is chosen by the monetary authority depending on deviations in output from potential and inflation from the target. So once the monetary authority has decided on the appropriate real interest rate, it will change the money supply, and thereby shift the LM curve to meet that target. Keep in mind this all works because prices are more or less fixed in the short-run so that real balances change with the money supply alteration.

Extension to the Open Economy

This basic set-up needs to be extended slightly to consider an open economy. The only change that occurs in this process is a variation in the composition of the IS curve. The IS curve was previously derived by equating planned to actual expenditures, where planned expenditures were the sum of consumption, investment, and government spending. Planned expenditures now also include net exports, or the value of exports less imports. Often, total trade flows are disaggregated into fuel and non-fuel goods and services. For most countries trade in fuel goods and services are relatively small, so the bulk of trade is in non-fuel goods.

Exports of non-fuel goods (x) can follow a standard error-correction framework (the c are estimated

coefficients, lowercase represents logs, and Δ the first-difference):

$$\Delta x_t = \Delta wt_t - c_1 cu_t - c_2 \Delta wcr_t - c_3[x_{t-1} - wt_{t-1} - c_4 trx_{t-1}] \tag{63}$$

According to this equation the percent change in exports depends directly on the change in world trade (wt), capacity utilization (cu) in the exporting country, and relative unit labor costs of the exporting country (wcr). In the long-run exports stay in-line with world trade and a time trend (trx) which captures changes in the share of a countries trade with the rest of the world. Imports of non-fuel goods (m) have a similar structure:

$$\Delta m_t = d_1 \Delta tfe_t + d_2 \Delta wcr_t - d_3[m_{t-1} - tfe_{t-1} - d_4 wcr_{t-1} - d_5 cu_{t-1}] \tag{64}$$

Imports directly depend on total final expenditures (tfe) in the importing country, or actual expenditures in the notation of the theoretical model, and relative unit labor costs. In the long-run imports move in relation to total final expenditures, unit labor costs, and capacity utilization.

Taken together, these equations show that net exports are determined in the short-run by demand (or expenditures). Although there is supply representation via unit labor costs in the short-run, these respond to demand conditions as will be outlined below. The longer-run contains elements of both supply and demand. The IS curve will now shift when foreign demand changes relative to domestic demand as well. For example, if there is higher foreign demand and domestic output stays the same, this raises exports, shifting out the IS curve.

The IS curve can now shift in response to changes in the real exchange rate as well. This would alter world trade, impacting exports, and possibly unit labor costs, altering both exports and imports. In some models the real exchange rate, how many foreign goods can be purchased with one domestic good, depends on the relative real interest rates between countries. A rise in the foreign real interest rate, for example, makes it more attractive to invest abroad, increasing net capital outflow, the corollary of which is a rise in net exports. Nominal exchange rates will often be derived based on the real exchange rate movements.

Full Employment Schedule

The final piece of the model is not standard, but helps the model to integrate some supply-side features. Until this point, the entire derivation of both the IS and LM curves has been demand driven. We now add the FE line, which shows equilibrium in the labor market. The full closed-economy $IS - LM$ model with an FE line is shown in Figure 5. This is the full-employment level of employment, and depends on the current levels of capital and productivity. Because it does not vary with the real interest rate, this is a vertical line in the plot between the real interest rate and output.

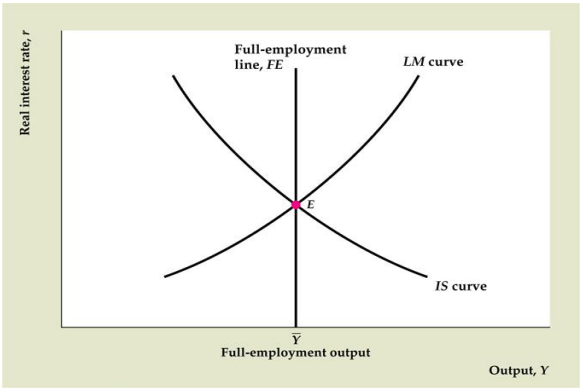


Figure 5: The $IS - LM$ Model With an FE line, from Abel et al. (2007)

Aggregate Demand

Another way to look at the $IS - LM$ relationship is using an aggregate demand curve. This schedule summarizes the relationship between the price level and output as shown in Figure 6, and can be derived from the $IS - LM$ model.

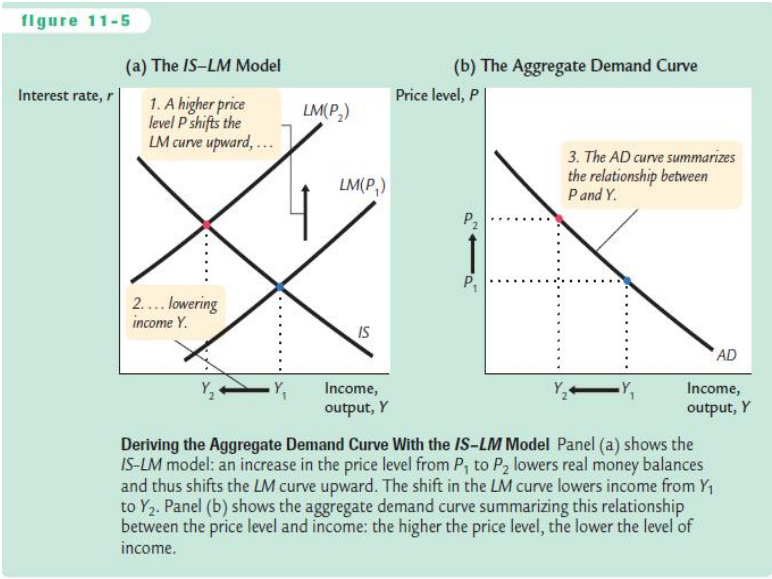


Figure 6: Derivation of the Aggregate Demand Schedule, from Mankiw (2002)

Suppose the price level rises. This will shift the supply of real money balances to the left, which indicates that for any given level of income the real interest rate is higher. As shown in Figure 6, this means the LM curve must shift up. The result is lower output, indicating an inverse relation between the price level and output as shown in panel (a) of Figure 6.

Supply

The short-run supply-side of the model works through both sticky wages and sticky prices. Beginning on the wage side, it is assumed that firms produce according to a production function which depends on both capital and labor. Capital is relatively fixed in the short-run, and as more labor is added additional output can be produced, but this occurs at a decreasing rate. An example of a production function with this shape is shown in Figure 7.

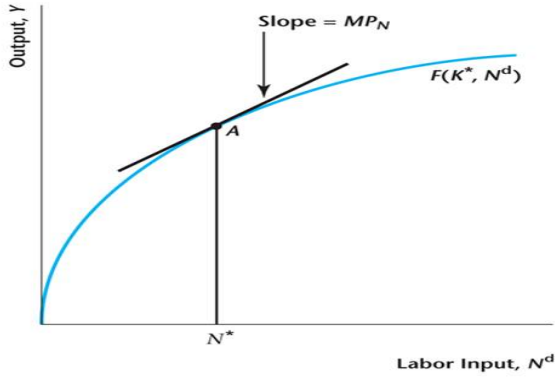


Figure 7: A Standard Production Function, from Williamson (2011)

A profit-maximizing firm will produce where the marginal product of labor is equal to the real wage rate. The marginal product of labor is the slope of the production function shown in Figure 7. These various slopes are traced out in Figure 8 for different levels of employment. Because everywhere along this line gives the value of the marginal product of labor, and a profit-maximizing firm equates the real wage rate to this marginal product at an optimum, Figure 8 is also the labor demand curve of the firm.

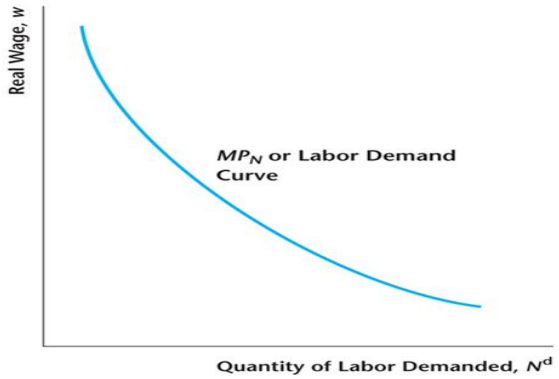


Figure 8: The Labor Demand Curve, from Williamson (2011)

The firm will hire workers until the marginal product equals the real wage rate, and this point can be read off the labor demand curve. Many models assume that labor demand will determine the amount of work in the market in the short-run. The easiest way to think about this is that there is a vertical

labor supply curve at the natural rate of unemployment. The next step is to assume that nominal wages are fixed at some level. Assume for now that real wages are not fixed, and can change with the price level.

This allows derivation of a relationship between the price level and output on the supply side. Suppose the price level rises with fixed nominal wages. This will reduce real wages, and the labor demand curve shows that lower real wages lead to higher demand for labor, which results in more production. This construction is shown in Figure 9.

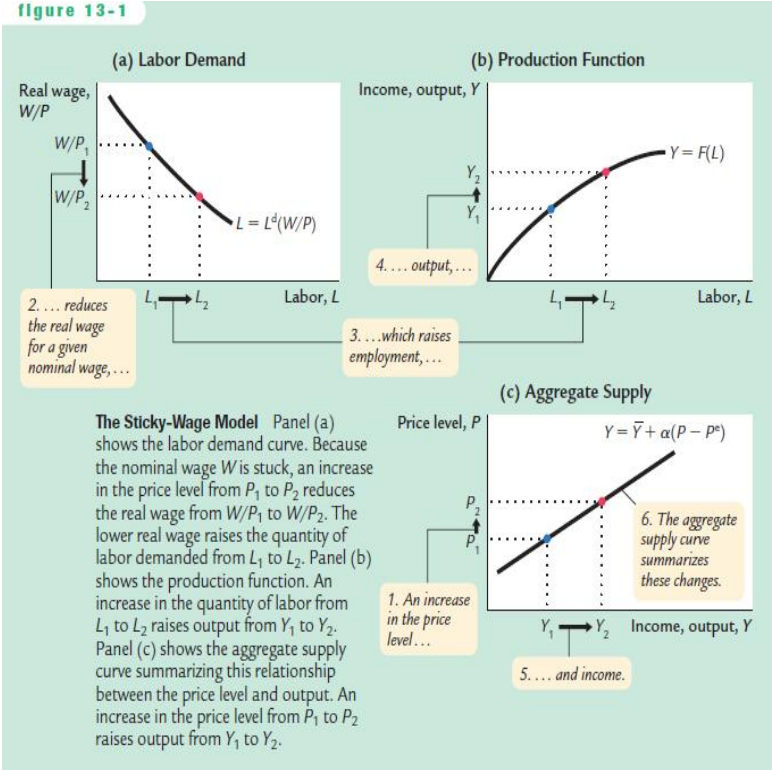


Figure 9: Derivation of the Aggregate Supply Curve, from Mankiw (2002)

The primary result of this sticky wage assumption is that there is a positive relationship between the price level and output. Intuitively, the sticky wage means that firm costs fall as prices rise, leading to higher profits, which entices them to hire more workers for additional production (and profits). This also allows the derivation of a relationship between wages and prices which is based on the slope of the labor demand curve.

Some macroeconometric models also assume that prices do not change much in the short-run, which can be due to market power or the costs of changing prices. This relationship imposes a close-to-horizontal shape to the aggregate supply curve. The combination of nominal wage and price

stickiness implies that real wages are sticky.[7] In the face of falling demand, firms need to either reduce their labor input, reduce real wages, or both. Because of the relatively fixed real wage, firms choose to reduce labor input, generating involuntary unemployment. This involuntary unemployment is in relation to the natural rate of employment. Think of this as a shift in the labor demand curve instead of a move along the curve. It must be a shift because moving along the curve means that the real wage is able to adjust. The result reinforces the shape of the upward sloping aggregate supply curve with involuntary unemployment generated by the combination of sticky prices and wages.

The derivation of aggregate supply in the long-run will be outlined in detail below. It is sometimes useful to use as a limiting case a vertical aggregate supply curve to depict long-run supply. The fact that it is vertical means that in the long-run supply does not depend upon the price level. Rather, it depends on the productive capacity of the economy.

Aggregate Demand and Aggregate Supply

The easiest way to understand the model in the short-run is through the $IS - LM$ set-up outlined above. However, there are some important concepts which can be illustrated by using the aggregate demand/aggregate supply framework shown in Figure 10

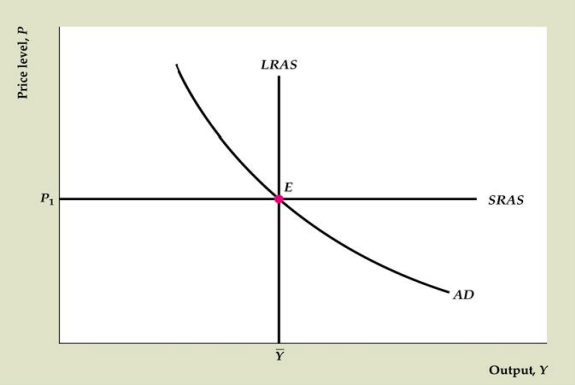


Figure 10: Aggregate Demand and Aggregate Supply in Equilibrium, from Mankiw (2002)

Most macroeconometric models imply a relationship between output and inflation in the short-run. That is, by increasing output (through government spending for example) a policymaker could get all of the benefits (such as reduced unemployment) at the cost of higher inflation.[8] This is straightforward to see with the AD/AS set-up. A rise in government spending shifts out the aggregate demand curve, leading to both higher output and a higher price level in the short-run, as shown in Figure 11.

[7] The real wage (w) is the nominal wage (W) deflated by the general price level (P), or $w = \frac{W}{P}$.

[8] This Phillips Curve relationship seems to hold over some periods in the data and not others for the U.S., but it is generally believed such a permanent trade-off does not hold.

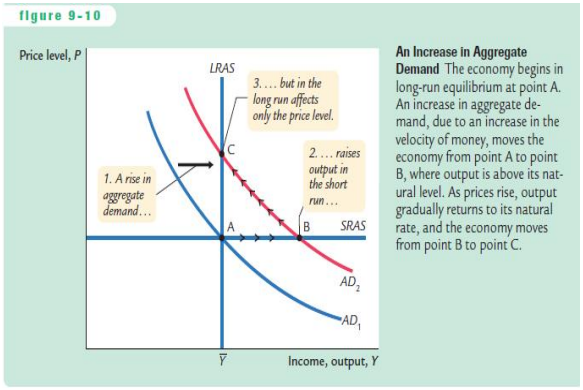


Figure 11: Change in Aggregate Demand, from Mankiw (2002)

But this tradeoff only holds in the short-run due to the long-run aggregate supply curve. Because increased spending does not change the productive capacity of the economy, there is no response in either short or long-run aggregate supply. Eventually the level of output must return back to the amount implied by the productive capacity of the economy, and this is also shown in Figure 11. The long-run result is a higher price level with no additional output above that implied by the economy's productive capacity.

The Theory of Macroeconometric Models: Long-Run

The long-run properties of popular macroeconometric models are based on the neo-classical growth model of Solow and Swan. This section outlines the basic theory and illustrates some ways in which the theory can be used in macroeconometric models.

The Solow-Swan Model

The Solow-Swan theory isolates the key factors of economic growth in the long-run based on a one-sector production function. The key lessons that emerge are the importance of technology and savings in generating consistent and sustainable economic growth. In what follows the basic theory is outlined by setting out the key equations, and then manipulating them to yield the primary equation which summarizes model dynamics. In the steady state, this equation is then shown graphically, and some simple experiments are conducted.

Equations and Assumptions

The Solow-Swan model assumes a constant and exogenous population growth rate:

$$N_{t+1} = (1+n)N_t \tag{65}$$

N is the population, and it grows at the constant rate of n each period. This equation can be re-written as: $\frac{N_{t+1}}{N_t} = 1 + n$. In aggregate all of the consumers in the economy can either consume (C_t) or save (S_t) aggregate income (Y_t):

$$Y_t = C_t + S_t \tag{66}$$

Because this is a closed economy, savings must equal investment by definition $(S_t = I_t)$, which will be used below. Importantly, it is also assumed that consumers as a whole save a fraction of aggregate income, or:

$$C_t = (1 - s)Y_t \tag{67}$$

In this equation s is a fixed savings rate for the aggregate economy. This is where the Solow-Swan model diverges from general equilibrium models: there is no endogenous savings/consumption decision. Because of this, it is not necessary to model the preferences of the consumers, which is why the model is solved in aggregate terms. The fixed savings rate implies that savings is a fixed fraction of aggregate income:

$$S_t = sY_t \tag{68}$$

This completes the description of the consumers in the model. There is no explicit representation of the firm, but the model does have an aggregate production function:

$$Y_t = Z_t F(K_t, N_t) \tag{69}$$

As is standard, production depends on capital and labor, and can be altered by technological progress. Capital also follows the standard accumulation equation:

$$K_{t+1} = K_t(1 - \delta) + I_t \tag{70}$$

Capital in the next period is investment plus any undepreciated capital. At this point, achieving consistency in the model requires assuming that all markets clear. There are two markets, one for consumption of current goods and a capital market (there is also a labor market, but this clear by assumption in the model). The market clearing condition for the capital market was stated above, that savings equals investment. In the current goods market, the clearing condition states that any

production must be used for consumption or investment:

$$Y_t = C_t + I_t \tag{71}$$

This is the equation which can be manipulated to generate diagrams. First, substitute out both consumption and investment using two of the equations above to give:

$$Y_t = (1-s)Y_t + K_{t+1} - K_t(1-\delta) \tag{72}$$

Next, substitute out income and insert the production function:

$$Y_t = (1-s)Z_t F(K_t, N_t) + K_{t+1} - K_t(1-\delta) \tag{73}$$

This expression can be rearranged to put it in terms of capital next period:

$$K_{t+1} = sZ_t F(K_t, N_t) + K_t(1-\delta) \tag{74}$$

There is one more step which is standard before obtaining a graphical representation, which is to put the variables in per-capita terms. This is done because the model does not have an explicit representation of welfare, and income per capita is used as a proxy. To put in this form, divide both sides by N_t:

$$\frac{K_{t+1}}{N_t} = \frac{sZ_t F(K_t, N_t) + K_t(1-\delta)}{N_t} \tag{75}$$

Notice that the left-hand side has $t+1$ over t, which gives inconsistent dimensions. To get around this, multiply the left-hand side by 1, which is the same as $\frac{N_{t+1}}{N_{t+1}}$:

$$\frac{K_{t+1}}{N_{t+1}}\frac{N_{t+1}}{N_t} = \frac{sZ_t F(K_t, N_t) + K_t(1-\delta)}{N_t} \tag{76}$$

Now use the fact that $\frac{N_{t+1}}{N_t} = 1+n$ and substitute in the left-hand side. Also, because the production function has constant returns to scale, the N_t can be factored out of this equation when in per-capita terms. The final equation, with lower-case reflecting per-capita variables is:

$$k_{t+1} = \frac{sZ_t f(k_t) + k_t(1-\delta)}{(1+n)} \tag{77}$$

This is the key equation in the Solow-Swan model. It describes the evolution of capital per person in

the economy. Capital per-capita is the most important variable, because as shown in the equation, it determines output per-capita. So higher capital per person means a higher standard of living in the model.

Graphical Representation

A good place to begin analysis is to plot capital between two adjacent periods and derive the steady state level of capital, as in Figure 12.

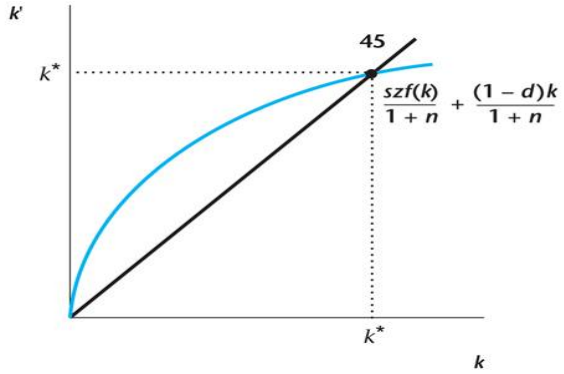


Figure 12: Steady State Capital in the Closed Economy Solow-Swan Model, from Williamson (2011)

The steady state level of capital stock is that which the economy tends to in the long-run. In this model, if the current capital stock is below the steady state level (k^*), investment exceeds capital depreciation so that the capital stock grows. To see this on the diagram, pick any point k_t below k^*. Project this point onto the figure, and the value of capital next period is above the 45 degree line. Because the 45 degree line gives all of the points where $k_t = k_{t+1}$, any points above this line indicate the capital stock is large in the next period.

This graph can be used to experiment with the model by changing various inputs and assessing their impacts on the steady state level of capital stock. However, while this representation makes the steady state level of capital clear, it is not the best way experiment with the model. For a better representation, use equation (77) but assume the model is at steady state, so that $k_t = k_{t+1} = k^*$. This gives:

$$szf(k^*) = (n + d)k^* \tag{78}$$

At the steady state this relationship must hold, so that per-capita investment (the left-hand side) is equal to the steady state level of capital when population growth and depreciation are accounted for. This relationship can be plotted, with steady state level of capital on the horizontal and either the left or right-hand side expressions on the vertical (by definition they are equal). This is shown in Figure 13.

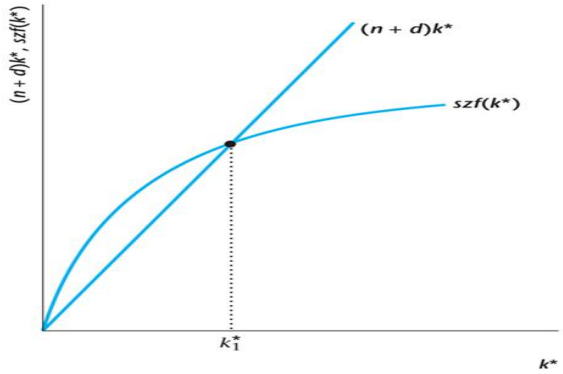


Figure 13: An Alternative Representation of Steady State Capital in the Closed Economy Solow-Swan Model, from Williamson (2011)

The shape of the production function gives the left-hand side its shape, while the right-hand side is linear as represented in the figure.

Shocks

We are now in a position to change any of the exogenous variables (n, d, s, z) to see the impact they have on the steady-state level of capital, and therefore long-run living standards. Figure 14 shows the impact of an increase in the savings rate.

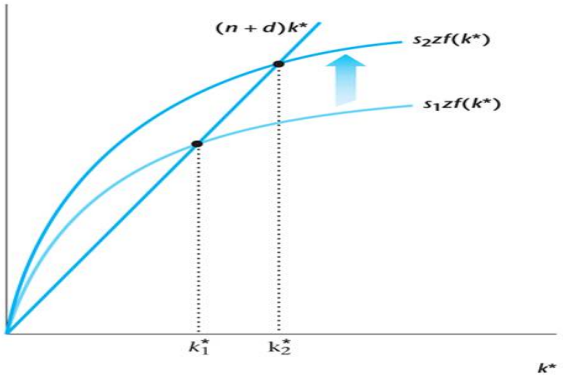


Figure 14: An Increase in the Savings Rate in the Closed Economy Solow-Swan Model, from Williamson (2011)

This increase will shift up the curve, and leads to a higher level of capital per person, which leads to higher GDP per person in the long-run. Thus increased savings should increase standards of living. This happens because consumers give up some consumption today to build capital, which has a higher payout in the future. While this is an intuitive result, there is only so high the savings rate can rise, so there are limits to this type of growth. The next experiment increases either the population growth rate or the depreciation rate of capital (or both) as in Figure 15.

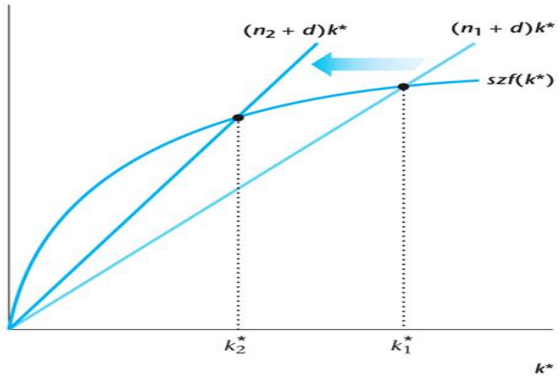


Figure 15: An Increase in the Population Growth Rate or Depreciation Rate in the Closed Economy Solow-Swan Model, from Williamson (2011)

This shifts up the other curve and results in a lower level of capital in the long-run. This result is also intuitive, in that if there are more people or capital wears out faster, there is less capital to go around per person. This is what is reflected in the graph. The final experiment gives the most important lesson from the Solow model, as it increases TFP as in Figure 16.

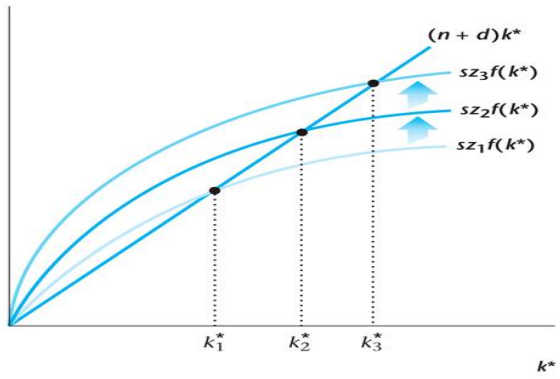


Figure 16: An Increase in Total Factor Productivity in the Closed Economy Solow-Swan Model, from Williamson (2011)

Increases in total factor productivity continually increase living standards through more capital per person. The strength of this type of growth is that it is potentially unlimited.

Other Variables

The remainder of variables in many models adjust to their long-run levels based on the assumed growth rates of productivity, population growth, and savings as outlined in the model above.

References

Abel, Andrew B., Ben S. Bernanke, and Dean Croushare, *Macroeconomics*, 6th ed., Addison Wesley, 2007.

Cochrane, John H., "Time Series for Macroeconomics and Finance," Lecture Notes 2005. Available at: `http://faculty.chicagobooth.edu/john.cochrane/research/papers/time_series_book.pdf`.

Enders, Walter, *Applied Econometric Time Series*, 3rd ed., Wiley, 2010.

Fair, Ray C., *Estimating How the Macroeconomy Works*, 1st ed., Harvard University Press, 2004.

_ , "The U.S. Model Workbook," Mimeo 2012. Available at: `http://fairmodel.econ.yale.edu/wrkbook/xazwrk.pdf`.

Kennedy, Peter, *A Guide to Econometrics*, 6th ed., Wiley-Blackwell, 2008.

Mankiw, Greg, *Macroeconomics*, 5th ed., Worth, 2002.

Nelson, Charles R. and Charles I. Plosser, "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, 1982, *10* (2), 139–162.

Stock, James H. and Mark W. Watson, *Introduction to Econometrics*, 2 ed., Pearson, 2007.

Taylor, John B., *Macroeconomic Policy in a World Economy: From Econometric Design to Practical Operation*, online ed., W.W. Norton, 1993. Available at: `http://www.stanford.edu/~johntayl/MacroPolicyWorld.htm`.

Williamson, Stephen D., *Macroeconomics*, 4th ed., Pearson, 2011.