Advanced Topics

Graduate Macro II, Spring 2010
The University of Notre Dame
Professor Sims

In this document I (briefly) lay out a few "advanced" topics that we did not have time to cover in the course. You should be vaguely familiar with these topics for the comprehensive exam.

1 Beyond Calibration: Model Estimation

As noted in class, calibration involves picking model parameters in order for the model to match certain low frequency (i.e. long run) features of the data. One might argue that a model's parameters should be chosen so as to match *all* features of the data, not just long run averages. Furthermore, there are many parameters that might not have anything to do with long run averages, and thus cannot be identified from a simple calibration exercise. Finally, we might be interested in knowing something about how certain we are about particular parameter values – i.e. interested in standard errors.

Formal model estimation takes the above issues seriously and goes beyond calibration. I will discuss, very briefly, three different approaches to estimation – method of moments, simulated method of moments/indirect inference, and maximum likelihood. For all of these, I will use the notation that Θ is a $q \times 1$ vector of the model's parameters.

An economic model has to first be solved in order to estimate its parameters. Typically we think about linearization as the solution to DSGE models, but that need not be the case. Once the model is solved, it is capable of producing data given parameter values. We can then pick parameters values that make the solved model "fit" the actual data in some sense to be defined below.

A good textbook reference if you're interested is Dejong and Dave, *Structural Macroeconomics*. A good shorter paper-length introduction is by Franceso Ruge-Murcia: http://www.cireq.umontreal.ca/personnel/ruge_methods.pdf).

1.1 Generalized Method of Moments

Loosely speaking, I'll refer to a "moment" as a statistic that can be observed in the data and that can be generated by a model.[1] So, for example, a mean is a moment, as is a variance or covariance. Formally, means are called first moments and variances are called second moments. I'll also use the terminology to describe other statistics of interests (like regression coefficients) which may not be moments at all in a statistical sense.

[1] Formally, a moment is the unconditional expectation of a random variable raised to a power – i.e. $E(x^s)$ is the s^{th} moment of the random variable x.

Method of moments estimation involves the researcher pre-selecting various moments of interest in the data. Given a model with stochastic disturbances which is treated as the data generating process of the real world data, the model's parameters are then chosen so as to make the model's moments as close as possible to the moments observed in the data.

Formally, let m^* be a $k \times 1$ vector of moments which you want to be able to match. These should be things which have a clear population counterpart in the model (i.e. like an unconditional mean or standard deviation). Let $m(\Theta)$ be the vector of those same moments from the model evaluated at a given parameter vector, Θ. Let $\mathbf{W}$ be a $k \times k$ matrix. The GMM estimate of the model's parameters is:

$$\Theta^* = \arg\min \qquad (m(\Theta) - m^*)' \mathbf{W} (m(\Theta) - m^*)$$

Essentially, the parameters of the model are chosen so as to minimize the weighted sum of squared deviations between model and data moments. Theory does not guide a choice of $\mathbf{W}$ – choosing an identity matrix, for example, would simply pick the parameters to minimize the sum of squared deviations between model and data moments. Statistics may have something to say about the choice of $\mathbf{W}$, however. The empirical moments, m^*, since they are taken from a finite sample, are subject to statistical uncertainty. It therefore might make sense to place less weight on those empirical targets which are estimated imprecisely. That's exactly what the choice of the "optimal weighting matrix" does – essentially, the optimal weighting matrix is the inverse of the variance-covariance matrix of empirical targets. The higher the variance of a particular target, the less weight it gets in the estimation. The choice of $\mathbf{W}$ will not affect consistency of estimates, but it will affect the magnitude of the standard errors (i.e. it's about efficiency).

Since there is statistical uncertainty in the empirical targets, there will also be statistical uncertainty in your estimate of Θ. As such, you can construct standard errors for the estimates given above. I will not go into details here. Basically, the calculation of standard errors will depend on how much variation you have in the data (i.e. the variance of the moments), and how informative your model is about those moments (i.e. the Jacobian matrix . . . how much do the model moments change when the primitive parameters change).

The researcher has some leeway in terms of what moments to target and how many of them to target. In order for the model to be identified, you need at least as many empirical targets as you have parameters of interest; i.e. you need $k \geq q$. When this holds with equality we say that the model is just identified. When it holds with inequality we say that the model is overidentified. Overidentification is typically considered a good thing – overidentification allows us to simultaneously estimate and test a model. The essential intuition is to use a subset of the moments to estimate the parameters, and then see whether or not the model fits the other moments (i.e. the ones not used in estimation). If it does not, we can "reject" the model as the true data generating process.

An example would be as follows. Take a simple real business cycle model. Take the following as empirical targets: the variances of output and consumption, the covariance between output and consumption, and the autocovariance of both output and consumption (i.e. five moments). Estimate the persistence of technology shocks, the standard deviation of technology shocks, the discount factor, and capital's share.

1.2 Simulation Based Estimators

Both the simulated method of moments and indirect inference are simulation based estimators. The estimation of parameters is conceptually identical to GMM. The main difference is that, instead of using population moments from the model, $m(\Theta)$, we simulate a number of different data sets from the model at the given parameter vector. Then we calculate our "moments" from the simulated data, and compare it to the moments observed in the actual data.

Since the estimator is conceptually identical to what is above, I do not repeat the optimization problem here. Calculation of the optimal weighting matrix and standard errors is a bit more involved, since there is now sampling error in the calculation of model moments. Using simulated method of moments would be identical to the example given above in the GMM section, but the model moments would be done from a simulation. What distinguishes indirect inference from simulation method of moments? Indirect inference typically encompasses looking at "non-moments", in the formal sense, from the model and comparing them to the data. For example, suppose you run a regression of consumption on output in the data, and get parameter estimate $\widehat{\lambda}$. This parameter may or may not have any structural interpretation in the model; even if it does, the conditions under which OLS will yield a consistent estimate may not be satisfied. But it's still a statistic of interest which your model ought to be able to match. So you could simulate data from your model and run the regression on the simulated data. You would then iterate on guess of the vector of structural model parameters, Θ, until the estimate of $\widehat{\lambda}$ you get in the model simulations is as close as possible to what you get in the actual data.

A natural question is why one would prefer simulation based estimators over conventional GMM, which uses population moments. The simple answer is that it may difficult (or impossible) to find the population "moments" of interest in the model. Simulation allows you to overcome this difficulty. There's also a sense in which simulation is in fact the right thing to do. You take the model as the true data generating process. The moments we observe in the data come from a finite draw of stochastic disturbances in the model. It seems natural that we should calculate moments from finite draws of the model in the estimation stage rather than using population moments.

1.3 Maximum Likelihood

Moment based estimators are called "limited information" methods for estimating parameters. The name is appropriate – a limited/finite set of information/moments are used to estimate the model's parameters. "Full information" are called maximum likelihood estimators. Maximum likelihood views the DSGE model as a probabilistic model; it then picks parameters of the model that make the observed data the "most likely" to have been generated from the parameterized model. What I present below is a "classical" approach to MLE.

In order to pick the parameters of the model to maximize the likelihood of having observed a particular sequence of variances, we need to first figure out the likelihood. Recall that a linearized DSGE model can be written state space notation:

$$x_t = Cs_{t-1} + De_t$$

C and D are functions of the deep parameters of the model. s denotes the state and x includes both the state and jump/non-predetermined variables. e is the shock vector; maximum likelihood requires a distributional assumption (normality) on these errors, whereas GMM does not, in general. Let $x^{t-1} = \{x_j\}_{j=1}^{t-1}$ denote the realization of the variables x from the beginning of time (period 1) up until the period before the present. Let $L(x_t \mid x^{t-1})$ be the probability of observing a particular x_t conditional on its past realizations. These conditional likelihoods are independent across time (as long as the shocks are iid), so the likelihood associated with observing an entire sequence of data is the product of the conditional likelihoods each period:

$$L(x) = \prod_{t=1}^{T} L(x_t \mid x^{t-1})$$

For an initial condition, it is typically assumed that $L(x_1 \mid x^0)$ is simply the unconditional likelihood. Using the state space structure above, the optimal one period ahead forecast of the variables of the model is:

$$\widehat{x}_t = Cs_{t-1}$$

The discrepancy between actual and observed values is the "residual":

$$D\widehat{e}_t = x_t - Cs_{t-1}$$

Thus, the probability of a particular realization depends on the probability distribution of $\widehat{e}_t$. Given normality of shocks, you can compute the probability of a particular realization of x. You can then go through and do this for the entire sequence of data. Because taking a natural log is a monotonic transformation, then we can write the problem as:

$$\Theta^* = \arg\max \qquad \ln L(x) = \sum_{t=1}^{T} \ln L(x_t \mid x^{t-1})$$

Since the parameters of the state space representation are functions of the underlying parameters, the likelihood is a function of those parameters. We pick the parameters so to generate the parameter vector where the actual realization of data is "most likely".

There is also a Bayesian interpretation of maximum likelihood that is frequently used in the literature. I'm not a Bayesian so don't take any of this too seriously. Loosely speaking, classical statistics views parameters as fixed and data as exhibiting some randomness. Bayesian methodology flips this – the data is given and the parameters have some randomness. Viewing the data as fixed, the researcher can construct a posterior distribution of parameters. The means or modes of the distribution are interpreted as the point estimates; the second moments of the distribution are the standard errors. One potential advantage is that the researcher can use a priori information to be impose "priors" on the distribution of structural parameters.

1.4 Comparison

The so-called "likelihood principle" says that all of the information about a sample of data is contained in the likelihood function. Thus, maximum likelihood is called "full information" estimation, as opposed to the limited information structure of moment based estimation. In principle, using more information is better than using less, so a researcher should prefer maximum likelihood.

I don't completely agree with this (my personal opinion). Why? I have little or no intuition for maximum likelihood. Moments based estimation is very intuitive to me – I have empirical targets I want my model to match, and I can see how well (or how poorly) my model does in matching those moments. As such, there is a clear way for me to evaluate the performance of my model and the plausibility of it being correct and useful for various forms of analysis. That doesn't necessarily emerge from maximum likelihood. Another reason to prefer moment based estimation (in particular simulation based methods) is that sometimes the likelihood doesn't exist or is really hard to form. Because computing power is now relatively cheap, there are reasons to go this route. One reason to prefer likelihood based estimation is that poor identification of parameters can be partially overcome by using a priori information in imposing priors.

2 Informational Frictions

There has been a lot of research (including some of my own) on incorporating informational frictions into DSGE models. Sometimes this involves departures from rational expectations (e.g. adaptive expectations or least squares learning). More often it involves agents (or some subset of agents) not being able to perfectly observe all relevant fundamentals/states. This imperfect information produces interesting things – agents might respond to non-fundamental noise, and noisy observations might alter the response to some structural shocks.

Much of this literature makes use of the Kalman filter, which is a way to optimally forecast unobserved variables given some observables. I explain the Kalman filter in detail in another document. Basically, the Kalman filter assumes that we have a state space system in which there is imperfect observation. The true state evolves according to:

$$x_t = Cx_{t-1} + e_t$$

The researcher only observes y_t, which is linearly related to x_t as:

$$y_t = Dx_t + u_t$$

u_t is a vector of measurement error/noise. Both e_t and u_t are iid with known, finite variances. The agents of the model will need an estimate of x_t to solve their problems. Suppose they take a linear filter as follows:

$$\widehat{x}_{t|t} = \widehat{x}_{t|t-1} + K(y_t - \widehat{y}_{t|t-1})$$

The idea of the filter is that your best guess of the current position of the state condition on the current information set ($\widehat{x}_{t|t}$) is equal to your best guess of what the current state

would be conditional on the previous period's information set plus an "adjustment factor", K, times the the forecast error in the observable (i.e. what you observed less what you thought you would observe today in the previous period). The forecast error in the observable comes about for two reasons – either there were true structural shocks, e_t, or there was noise in the past that caused you to come up with an erroneous estimate of the state. The goal is to find the K that minimizes the forecast error variance:

$$K^* = \arg\min \quad var(x_t - \widehat{x}_{t|t})$$

Let's take a simple example here; for the complete derivation, see my other notes. Suppose that the state space system is particularly simple:

$$\begin{aligned} x_t &= e_t \\ y_t &= x_t + u_t \end{aligned}$$

In other words, the state variable of interest is just white noise; you observe this variable with noise itself. The forecast error variance is:

$$var(x_t - \widehat{x}_{t|t}) = var(x_t - Ky_t) = var(e_t - K(e_t + u_t))$$

This follows from the fact that, since there is no time dependence in the model, your forecast of future ys will always be zero. We can work this out as follows:

$$\begin{aligned} var(x_t - \widehat{x}_{t|t}) &= E(e_t - K(e_t + u_t))^2 \\ &= E\left(e_t^2 - 2K(e_t^2 + e_t u_t) + K^2(e_t + u_t)^2\right) \\ &= \sigma_e^2 - 2K\sigma_e^2 + K^2\left(\sigma_e^2 + \sigma_u^2\right) \end{aligned}$$

The last line follows from the fact that disturbances are mean zero and uncorrelated as well as the linearity of the expectations operator. Taking the derivative with respect to K and setting it equal to zero yields:

$$\begin{aligned} -2\sigma_e^2 + 2K\left(\sigma_e^2 + \sigma_u^2\right) &= 0 \\ K &= \frac{\sigma_e^2}{\sigma_e^2 + \sigma_u^2} \end{aligned}$$

Essentially, the optimal Kalman gain (that's what we call K) is equal to the signal to noise ratio. If you observe the state with a lot of noise (σ_u^2 big), you basically don't ever update your perception of the underlying state.

The Kalman filter is not only useful when thinking about explicit informational problems, it's also necessary to use it when part of the state vector is hidden when constructing the likelihood function in MLE estimation of a model above.

3 Financial Frictions

Most of the models with which we've dealt have had perfect credit markets. By perfect credit markets we mean that agents can freely exchange securities, that there are no agency costs, no liquidity issues, etc.. Because the financial sector works perfectly, we typically ignore it completely in conventional business cycle models. That is obviously a large omission, particularly in light of what has happened over the last several years.

Fundamentally, the financial sector in general, and banks in particular, serve the purpose of matching savings with investment. The friction that the presence of banks resolve is that investment projects are typically very large and the savings of individuals is very small. So the bank pools savings from individuals, and in so doing can fund large, illiquid investment projects. If credit markets work perfectly, we can ignore this altogether in our models. But there are reasons to think that credit markets don't work perfectly.

One basic problem is that the funds of depositors get tied up in illiquid investment projects. If enough of the depositors decide that they want their money back, the bank won't be able to meet its obligations. It will fail, and the investment project will collapse, wreaking havoc on the real economy. These "bank runs" were one of the essential problems of the Great Depression. Something like deposit insurance/promises of bail outs should prevent these runs from happening in the first place.

But promises of bailouts leads itself to the problem of moral hazard – banks have incentives to take too much risk if they think that they will be bailed out. Furthermore, there is another layer of moral hazard – the principal-agent problem. The people who run banks and the people who make investment decisions are separate from those who supply funding (the depositors). For both of these reasons, we often see banks having very high leverage – leverage can be thought of as the ratio of debt to equity. Basically, because banks think they'll get bailed out (or because the managers don't have "skin in the game", banks can take on too much risk by getting very high leverage ratios. High leverage makes the bank very susceptible to even small economic shocks. This is basically the idea of the financial accelerator. The basic idea is that firms rely on external funding for large projects. If their balance sheet worsens for whatever reason, it's harder to get this external funding, amplifying the effects of the shock that worsened the balanced sheet in the first place.

Thinking about financial market frictions in DSGE models promises to be a really important area of research in the coming years. If you are interested, here's a brief bibliography of important papers in this area:

Diamond and Dybvig, "Bank Runs, Deposit Insurance, and Liquidity", *JPE*, 1983

Bernanke and Gertler, "Agency Costs, Net Worth, and Business Fluctuations", *AER*, 1989

Bernanke, Gertler, and Gilchrist, "The Financial Accelerator and the Flight to Quality", *ReStat*, 1996

Kiyotaki and Moore, "Credit Cycles", *JPE*, 1997

Fostel and Geanakoplos, "Leverage Cycles and the Anxious Economy", *AER*, 2008

4 Medium/Large Scale DSGE Models

The models that we have examined in class are relatively small with few parameters. As argued, this model began as a neoclassical growth model with variable labor and stochastic productivity shocks. We talked about a number of modifications of that model. Basically, the current generation of "medium/large scale" DSGE models takes that neoclassical backbone and adds lots of bells and whistles and multiple stochastic disturbances. The canonical models here are by Smets and Wouters (2003 and 2007, the former for Europe and the latter for the US) and Christiano, Eichenbaum, and Evans (2005). Basically, these models have the following bells and whistles: habit formation in consumption, preference shocks, adjustment costs to investment and/or capital, sticky wages and prices (typically the assumption wage-setting is Calvo, just as with price-setting), government spending shocks, distortionary tax shocks, and a central bank that follows a Taylor type interest rule.

The following are the stochastic shocks that are typically entertained:

1. Preference shock 1 – current utility matters more relative to future utility
2. Preference shock 2 – leisure gets more weight in current utility relative to consumption
3. Technology shock – standard AR(1) process for TFP. Sometimes there are both transitory and permanent technology shocks; sometimes there are also news shocks
4. Investment specific technology shock – normal models assume that output can be converted into either consumption or capital just as easily. In other words, the relative price of investment and capital is unity. We could instead specify the aggregate accounting identity as: $y_t = c_t + z_t I_t$. Shocks to z_t alter the relative price of investment to consumption.
5. Government spending shocks – just as normal, usually an AR(1).
6. Tax shock – to the distortionary tax rates on labor and capital income.
7. Monetary policy shock – shock to the Taylor rule
8. Cost push shock – the Phillips Curve

You can see that you're quickly getting to have a lot of shocks. For each shock there are typically two parameters to estimate – the autoregressive term and the variance. So if you have 7 shocks, that's already 14 parameters. Then you have typically another 10-20 or so parameters related to preferences and technology. All the sudden the model gets pretty cluttered. Typically there are 30 plus parameters in the model, which will then be estimated, usually be maximum likelihood. Then with the parameterized model in hand, policy type experiments can be run.

You might not be surprised to learn that this approach is often criticized. The medium/large scale approach sets out to explain the data well; to do so, it needs lots of bells and whistles, because the aggregate US economy is very complicated. This means that these researchers often resort to what could be called "ad hoc" specifications. The models are

over-parameterized; it isn't clear where identification of these parameters come from and it's difficult to understand clearly the mechanisms behind which shocks affect the economy. For a good critique of this approach, see Chari, Kehoe, and McGrattan (2008).