607
Overview and intro

Jon Faust

http://e105.org/e607

September 6, 2016

- **Goal of course: give you a start on doing reliable and useful empirical work in macro.**
- **Gaol of second year courses**
  - Get you involved in the discussion
  - You need to be prepared to keep up for the rest of your career
  - We ease into that self-guided learning in second year courese
- **Requirements**
  - Problem sets, 25 %
  - Brief presentations, 15%
  - Final, 60 %
- **Parts**
  - Technical stuff about probability and stats.
  - Numerical stuff about how to use the computer to answer questions you can't answer with pencil and paper.
  - Practical stuff about good econometric practice.

    The equivalent of good lab. technique in the physical sciences.
  - Some exposure to the ins and outs of macro data.
- **A few words about each part**
- **Technical stuff**

- The technical stuff will build on what you know and emphasize special wrinkles added by time series.
- Also PhD courses are meant to give you the tools to keep up with the field on your own

  you should be well on your way
- In lectures, I'll mainly provide guides to help you read and learn technical stuff on your own.

▶ **Numerical**

- Many of our problems are quite messy and not amenable to analytic solutions

  Or to analytic solutions that are simple enough to understand.
- Computer power has reached the point that we can learn much more and learn it much more cheaply (in time and in scarce brain cells consumed) with some sort of computer simulation than with purely analytic techniques.
- Further, numerical methods can provide us the ideas about where there may be important analytic results.

▶ **Econometric Hygiene**

- There is simply no way to sugar coat the fact that standards of care and good practice in applied macro are very low.

  e.g., Papers cannot be replicated. Simple programing errors drive results that have been trumpeted as having great practical import.
- We'll discuss ways to do better.

▶ **Data**

- Our models have simple items like income, consumption, investment, and government spending and satisfy:

$$Y = C + I + G$$

- The analogs of these items (and others such as interest rates, etc.) in practice don't correspond very closely to the theoretical constructs

  And, e.g., don't satisfy $Y = C + I + G$.
- Doing good work requires understanding the real world measurement system generating these items that we use as analogs to our theoretical constructs.
- Understanding this is a lifelong endeavor, but we'll get a start.

▶ **Course website: http://e105.org/e607**

▶ **Texts, etc.**

- No official text
- Website and lectures will refer to many texts and readings

  Many of them available on the web.

- As a general background, I like Bruce Hansen's text which is up on the web.
- On specifics, I'll also assign parts of Jonathan Wright's notes

▶ **Overview: theory and practice**

▶ **In practice,**

- It is hard to do useful econometrics without deep understanding of theory
- It is hard to do useful econometrics if you don't understand profound limits on the practical relevance of much theory.

▶ **If you learn one thing...**

- Theorists give us lots of tools that deliver efficiency, optimality, robustness; ideal and superlative properties.

  Wow. That sounds really great. I'm all in favor of ideal things.

- If you learn one thing in this course, it is to ask 'In what sense?' when you hear any wonderful sounding properties
- More specifically, applied workers, at least, should ask, 'In what *practically relevant* sense?'

▶ **We'll learn**

- We'll learn that what is often labelled robust is often less reliable than stuff that is not labelled robust.
- We'll learn that optimal speeds of convergence generally mean nothing.
- We'll learn that what is a 'large' sample in one sense may not be large in many relevant senses.
- In short, these theory-based terms are often used in misleading manners.

▶ **We'll also learn**

- The theory behind these terms, if applied astutedly helps immensely as we attempt to organize and understand the practical relevance of results.

▶ **Simple overview: Elemetary to advanced to modern advanced econometrics**

▶ **Elemetary econometrics in a nutshell**

- Under some strong assumptions including linearity of the model, the OLS estimator is
  - Properly centered (in a certain sense)
  - Efficient (in a certain sense)
  - Exactly normally distributed giving us access to a large family of standard tests with t, χ^2 and/or F distributions under the null

▶ **Advanced econometrics**

- Under very weak assumptions and allowing a general smooth, but perhaps nonlinear model, some OLS-like estimator is,
  - Properly centered (in a certain sense)
  - Efficient (in a certain sense)
  - Asymptotically normally distributed giving us access to a large family of standard tests with asymptotic t, χ^2 and/or F distributions under the null
- (Note: these estimators are in the CAN framework, Consistent, Asymptotically Normal

▶ **One additional factor in moving from elementary to advanced**

- When we go to the weak assumptions we take on one additional issue
- The (asymptotically) efficient OLS-like estimator may have drawbacks

  practical (such as computational complexity) or statisical (some sense of robustness)

- Thus, in going from elementary to advanced, we get a fork in the decision tree:

  Should we use the efficient estimator or some less efficient estimator that has other desirable properties?

▶ **Historical note**

- The groundwork for much of advanced econometrics was laid out at least by the 1930s, but the profession took a long time to nail it all down
- But pretty much done by the end of the last century

▶ **From advanced to modern advanced**

- Econometricians viewed the results of advanced econometrics as nearly magical.
- Drop lots of assumptions, but still proceed essentially as you did in the elementary case.
- We just change the answer to 'in what sense?' a bit, but folks seldom ask this question,

  . . . we effectively act as if its magic.

- But, as already hinted, we discovered
  - In many contexts, the asymptotic approximations are hideous in relevant sample sizes.
  - The particular nature of time series dependence in macro data gives rise to special problems with the approximations.

▶ **Modern advanced econometrics**

- Modern advanced econometrics uses various theory tools to understand and structure the breakdown of the conventional asymptotic results
- And to suggest better methods.

  Q: Better in what sense? A: More reliable in relevant sample sizes.

- Two key tools: functional central limit theorems and higher order asymptotics.

▶ **Remainder today: Prob. and stats.**

▶ **Prob. and stats.**

- All our analysis will be underpinned by a complete probability model.
- In the remainder today, we'll review stuff you should already know and use this review as an opportunity to introduce the notation that will be with us through the class.

▶ **Basics and notation**

- Economic study often starts with parameterized models in what I call 'economic problem form':

  decision problems, objective functions, contraints, etc.

- I'll call this M_θ for $\theta \in \Theta$,
- This form of the model is economically interpretable, but the implications for observables and latent variables are often quite opaque

▶ **Solving the model**

- We solve the model to get a data generating process, DGP_θ.
- The DGP is generally a set of stochastic relations, perhaps decision rules for agents, describing a law of motion for obervables and latent variables
- DGP is essentially a recipe for simulating the model.

  Still leaves the stochastic relations among the variables rather opaque

- The model and hence DGP will then *imply* a joint probability distribution for observables
- We will say $M_\theta \Rightarrow DGP_\theta \Rightarrow P_\theta, \theta \in \Theta$.

- In this course, P will be a probability measure for observables, Y, $(T \times K)$ and perhaps latent varibles Λ, $(T \times L)$ and an initial condition ξ (smallish unspecified dimension).

▶ **Aside:: ARMAs, state space**

- In linearized models, the DGP implied by the model is generally an ARMA,

  which will have a state-space representation

- We'll learn these notions shortly if you don't know them

▶ **Aside:: Microfounded**

- Work starting at M_θ where this describes optimization problems for all included problems is called microfounded
- Sometimes we just posit DGP_θ or P_θ, skipping the elemental description in economic terms

  That is, we simply posit a family of stochastic behavior without deriving from a economic problem primitives.

- The distinction here is between work that is called 'microfounded' or 'not microfounded'
- Work that is not microfounded is sometimes called *ad hoc*

  if someone accuses you of being *ad hoc*, they are probably attempting to politely tell you that you and your work are worthless

▶ **Aside::**

- Some reality: The distinction between microfounded and not microfounded is usually ill-defined.
- The distinction is used in foolish ways by the small minded and those with some agenda other than advancing the science.

  (Not that I have a strong opinion on the matter.)

- You'll have to find your own way here.

▶ **Fortunately, in this class**

- You learn your M_θs in economics classes.
- In econometrics class, we mainly learn tools for analyzing the implied P_θs
- Thus, the microfoundations debate won't concern us much

▶ **Some technical details**

- A complete probability model has three elements, a probability measure, a sample space (the set of all atomic outcomes), and a σ-algebra of measureable subsets of the sample space defining measurable events.
- In this course, we generally won't need to be specific about the sample space or sigma-algebra.
- At times, however, it will be useful to say that $Y \in \mathcal{Y}$, where $\mathcal{Y}$ can generally be taken to be $\mathcal{R}^{(T \times K)}$

▶ **Distribution function and density**

- Given P_θ, I'll often write, F_θ for the associated distribution function and f_θ for the associated density.
- We will regularly use the conditional-marginal factorization of joint densities:

$$f(a, b) = f(a|b)f(b)$$

- Compact notation: If we were being really explicit and f were a joint density for scalars a and b, the conditionals and marginals would be written

$$f_a(.), f_b(.), f_{a|b}(.|.)$$

  and so forth.
- I will often follow a standard practice (in sloppy notes, at least) of leaving out the subscripts on the fs, so that we have to deduce which version of the function we are using based on the arguments
- Specifically, we will often be conditioning on the past
- And when given the joint density, e.g., $f(x_t, x_{t-1})$, we will write the conditional as $f(x_t|x_{t-1})$, rather than $f_{t|t-1}(x_t|x_{t-1})$.

▶ **Aside:: Notation for random variables vs. realizations**

- Our variable Y will sometimes stand for a random variable and sometimes for a realization of a random variable.
- I could, e.g., put it in bold when its a realization and not bold when its not
- Instead, you'll usually have to figure this out in context.

  When it is important or subtle, I'll try to be explicit.
- Often when I want to emphasize that Y is a realization, I'll write Y^r.

▶ **Ok, that's the end of some basic notation.**

▶ **Bottom line**

- All observable implications of the models M_θ are captured in the implied probability distributions, P_θ.
- Macroeconometricians make a good living helping folks interpret P_θs implied by M_θs, because the P_θs with relevance to reality are really messy

  People who can do this well are in scarce supply.

▶ **Standard macroeconometrics**

- We assume that P_θ implies a distribution F_θ and density f_θ.
- $f_\theta(Y^r)$ interpreted as a function of θ for fixed Y^r is call the *likelihood function.*

  in Y^r, the r is for realized.

▶ **Big idea: We can think of much inference as summarizing the shape of the likelihood at Y^r for various θs**

▶ **Likelihood**

- I'll write the log of the likelihood function as follows:

$$\mathcal{L}(\theta|Y^r) = \ln(f_\theta(Y^r))$$

▶ **Frequentist inference: MLE**

▶ **MLE**

- Maximum likelihood estimator (MLE):

$$\hat{\theta}_{MLE} = \max_{\theta \in \Theta} f_\theta(Y^r)$$

- You should know standard regularity condions under which the MLE has standard desirable properties

  (consistent, asymptotically normal, attains the Cramer-Rao bound).

- The frequentist would love to report the whole likelihood, but in practice this is cumbersome, so the frequentist reports various summary measures.

▶ **Truth**

- To discuss accuracy of point or interval estimates, we need to bring up a thorny subject: truth.

- We'll generally write θ^* for truth.
- There, that wasn't so hard, was it.

▶ **Aside:: Truth, the thorny part**

- In macro, we often posit models that are obviously metaphors or crude approximations to reality.
- In such a context, truth can be ill-defined.
- E.g., we posit a representative agent.
- Then we talk about (or estimate) this agent's CRRA.
- But from mountains of evidence, folks' utility functions differ.
- And to get formal about it, those differences are not of a form that aggregate into a representative agent with CRRA utility.

  Hmmm, so what is truth here?

▶ **Ok, back to what maximum likelihood**

- Because we can't conveniently report the whole likelihood, we report summary measures.
- These include $\hat{\theta}_{MLE}$ and some estimate of precision

  (say, a confidence interval).

- A natural confidence interval is formed by the set of θs that are nearly as likely as $\hat{\theta}_{MLE}$.

  If the range of θs similarly likely to the MLE is large (in the relevant economic sense) the we can't be very confident of the value of θ. If the range is small, we are more confident about the value of θ.

- In well-behaved cases, the likelihood is approximately quadratic in a neighborhood of θ^* (truth).
- In this case, the curvature of the likelihood at $\hat{\theta}_{MLE}$ will give us a way to find the θs with P_θ close to $P_{\hat{\theta}_{MLE}}$
- Remember the relation between the Hessian of the likelihood and the variance-covariance matrix.

▶ **Slightly mathy**

- Define the score and the Hessian for the log-likelihood:

$$\begin{aligned} q(\tilde{\theta}; Y^r) &\equiv \partial \mathcal{L}/\partial\theta|_{\theta=\tilde{\theta}} \\ Q(\tilde{\theta}; Y^r) &\equiv \frac{\partial^2 \mathcal{L}}{\partial\theta\partial\theta'}|_{\theta=\tilde{\theta}} \end{aligned}$$

- The first order condition satisfied by the MLE is $q(\hat{\theta}_{MLE}; Y^r) = 0$ and taking an approximation gives
$$q(\hat{\theta}_{MLE}) \approx q(\theta^*) + Q(\theta^*)(\hat{\theta}_{MLE} - \theta^*)$$
- Rearranging and using that the LHS is zero:
$$(\hat{\theta}_{MLE} - \theta*) \approx -Q^{-1}(\theta^*)q(\theta^*)$$
- Multiply by $\sqrt{T}$:
$$\sqrt{T}(\hat{\theta}_{MLE} - \theta*) \approx -Q^{-1}(\theta^*) \times \sqrt{T}q(\theta^*)$$
- Under regularity conditions, the first term has a plim of $\mathcal{I}^{-1}$ and the second term is asymptotically distributed $N(0, \mathcal{I})$, where $\mathcal{I}$ is known as the information matrix.
- Thus,
$$\sqrt{T}(\hat{\theta}_{MLE} - \theta^*) \rightarrow_d N(0, \mathcal{I}^{-1})$$
- Thus, the dispersion of the MLE is intimately related to the curvature of the likelihood in the region of the MLE in large samples.

  I need the 'in large samples' part because the MLE is consistent and so in large large samples, the MLE tends to be very close to θ^*.
- This sort of reasoning should be familiar to you or you should review it.

  See, e.g., Hansen's text on Maximum Likelihood, especially appendix B.11.

▶ Bayesian inference

- Standard Bayesian inference relies on that complete probability model the frequentist analyzes
- The Bayesian adds a probability distribution over the parameter space, Θ, with density $p_0(\theta)$

  called a prior density over θ or simply a prior.
- For any $\theta_1, \theta_2 \in \Theta$, $p_0(\theta_1)/p_0(\theta_2)$ is taken to be the relative plausibility of the claim that θ_1 is the true vs. θ_2 is true.
- Start with:
  - a model, P_θ with likelihood f_θ
  - a prior p_0
- If we add a 'fresh' sample, Y^r, we can ask how we should change our views about the plausibility of various θs.
- That is, we'd like to update p_0 in light of Y^r.

- Bayes law tells us,

$$p_1(\theta) = p(\theta|Y^r) = \frac{f_\theta(Y^r)p_0(\theta)}{p(Y^r)}$$

  where $p(Y^r)$ is the unconditional distribution of the observed under the model and prior.

- Or

$$p_1(\theta) \propto f_\theta(Y^r)p_0(\theta)$$

  That $\propto$ means 'is proportional to'

- This latter tells you the shape of p_1, the posterior.

  To finish the job of evaluating the posterior just need to normalize to integrate to 1.

▶ **And so,**

- Bayesian would love to thoroughly report the posterior, but in practice is limited to various summaries of the posterior.
- The posterior itself is essentially a smoothed version of the likelihood
- For a 'flat' prior— $p_0(\theta) = \text{const}$ for all θ—the posterior has the same shape as the likelihood.

▶ **In practice,**

- The Bayesian reports, say, the max. (called the posterior mode in this case) and curvature at the mode of the posterior.
- And if the prior is pretty flat in a region of the max., this is the same as the max. and curvature at the max. of the likelihood.

▶ **So**

- Frequentist: Max. of likelihood and curvature at the max.
- Bayesian: Max. of smoothed likelihood and curvature at the max.

▶ **And in the limit as $T \to \infty$**

- In large samples and in well behaved problems, the prior becomes unimportant

  And Bayesian and frequentist inference give the same answer.

- Ok, this is true except in certain exceptional cases.

  The importance of these exceptions is hotly debated

▶ **Why all the friction?**

- Bayesians and frequentists are often strenuously at odds

but report approximately the same things

- The difference is in how they individually would interpret the results
- Of course, when you publish you have no control over how folks interpret the results
- So it is a little hard to get too excited.

▶ **Let's dip into time series**

▶ **Time series and the likelihood**

- With a sample of T iid random variables, the joint probability distribution is the product of the T univariate distributions:

$$f_\theta(Y) = \Pi_{t=1}^{T} f_\theta(y_t)$$

where a more complete notation would have $f_{\theta,T}(.)$ on the left and $f_{\theta,1}(.)$ on the right.

- The the log likelihood is a sum of identical things

$$\mathcal{L}(\theta|Y) = \sum_{t=1}^{T} \xi_t$$

where

$$\xi_t = \ln(f_\theta(y_t))$$

- For fixed θ and viewing Y as random, the ξs are iid when the ys are.

▶ **Having transformed the likelihood...**

- Having transformed the likelihood into a sum of iid things, we can now bring to bear a bunch of very simple statistical tools.
- Aside: Our ability to understand statistics, doesn't reach far beyond the behavior of normalized sums.
- To analyze sums, we can bring to bear lots of WLLN and CLT tools

▶ **Aside:: WLLN, CLT**

- Take $x_t, t = 1, \ldots, T$ with $Ex_t = \mu$, and define the sample mean, $\bar{x} = T^{-1} \sum x_t$
- WLLN: Weak law of large numbers gives additional conditions on the xs such that

$$\bar{x} \to_p \mu$$

- CLT: Central limit theorem gives initial conditions under which

$$\sqrt{T}(\bar{x} - \mu) \to_d N(0, \sigma^2)$$

for some $0 < \sigma^2 < \infty$.

▶ Time series

- In time series, observations are sequential in time and are related through time.
- So the joint distribution of T observations is not the simple product of the underlying marginal distributions of observations.

▶ Aside:: Reminder: conditional marginal factorization

- For any joint density, $f(a, b)$, we have

$$f(a, b) = f(b|a)f(a)$$

where $f(b|a)$ is called a conditional density and $f(a)$ is the marginal density for a.

▶ Time series and sequential factorization

- In time series, we often find it useful to condition on the past

$$f_\theta(y_t) = f_\theta(y_t|y_{t-})f_\theta(y_{t-})$$

where in this class y_{t-} will be all elements prior to t.

- We can write $f(y_2) = f(y_2|y_1)f(y_1)$ and $f(y_3) = f(y_3|y_{2-})f(y_{2-})$ and continue through the sample.
- The full liklihood is:

$$f(Y) = f(y_1)\Pi_{t=2}^{T} f(y_t|y_{t-})$$

- And the log likelihood is:

$$\mathcal{L}(\theta|Y) = \ln(f_\theta(y_1)) + \sum_{t=2}^{T} \xi_t$$

where

$$\xi_t = \ln(f_{\theta,t|t-}(y_t|y_{t-})$$

where for clarity, I have added the subscripts to f indicating which conditional distribution we are referring to.

- This looks nominally like the iid case, but we have two problems.
  - We have an initial condition (the bit with y_1)
  - The ξs need not be independent or identical. In particular, $f_{2|1}$ need not be the same as $f_{3|2,1}$ or $f_{4|3,2,1}$.

▶ The initial condition

- Often in time series we simply condition on the initial condition

  That is, given a sample, Y^τ, we condition our analysis on the value of y_1^τ.

- Also in large samples and in well-behaved cases, the sum will dominate the log-likelihood and the initial condition won't matter.
- However, in practical cases, the initial condition will sometimes matter.

  So we'll do some talking about that.

▶ **Non idd ξs**

- The fact that the ξs are not iid is a bigger problem.
- We have plenty of CLTs and WLLNs for heterogeous and dependent data, but we'll have to figure out how to use them cleverly
- And learn about when those theorems provide a useful approximation in relevant sample sizes

  this is really the hard part.

▶ **That's it for a quick summary of the course and a** review of basic results and notation.