607

From OLS to a large class of estimators in the CAN framework

Jon Faust

http://e105.org/e607

September 14, 2016

▶ **Summary**

- Our object is to get as painlessly as possible from OLS in the most elementary case to advanced time series
- Started with liner model with fixed regressors and Gaussian iid errors and the OLS estimator.

  found OLS unbiased, BLUE, and inference can be based on normal distribution.
- Went to OLS in much richer linear case—everything stochastic, heterogeneous and dependent.
- We found our way to consistent estimators where approximate inference can be based on an asymptotic normal distribution.

▶ **Almost there, but a couple more issues**

- 1. In the general case, there is no reason to suppose that linear models and OLS is all we need.

  Thus, we need to consider a much richer class of models and estimators
- 2. We have only gestured at efficiency in the advanced case

  mentioning that we may want to consider some GLS-like estimator
- This lecture takes up the first of these

  It's easiest discuss to efficiency once we have tackled the general estimators.

▶ **The key: When working in the CAN framework, a huge range of general models and estimators can be analyzed in a way that is nearly identical to the OLS case.**

▶ **Notation**

- Any variable with a bar over it is some sort of sample mean; anything with a hat over it is an estimator for the parameter under the hat.
- Items with hats and bars always implicitly have a sample size subscript, usually T We'll almost always leave this off for ease of reading.
- For the linear model, OLS discussion, we follow the conventions that the slope parameters are given by β, and the data are Y and X, the LHS and RHS variables, repsectively.
- In general models, there is no necessary distinction between LHS and RHS variables, so we will return to our general notation: θ is the parameter, Y is all the data, and $Y \sim P_\theta$.
- For the parameters, θ and β in this lecture, a star superscript marks the true value; subcripts a, b, etc. just indicate some arbitrary parameter value.
- We'll write $\partial f(\theta_0)/\partial\theta$ to mean the first derivative of $f(\theta)$ evaluated at θ_0.

▶ Motivation about general models

- We are discussing the CAN framework.
- Accept for a moment we have already taken care of the 'C' part. That is, we have a consistent estimator.
- And we only care about cases where the sample size is large enough for the asymptotic results to give a good approximation.
- This amounts to starting from the point that we have some θ^c for which $\theta^c \approx \theta^*$.

▶ Consistency and linear approximations

- But if the model is smooth, as we generally assume, and we have $\theta^c \approx \theta^*$, then we can take a linear approximation to the model and pretty much forget that the original is nonlinear.
- That is, if the economic model says $\Xi(Y;\theta) = 0$, we write

$$\Xi(Y;\theta^*) \approx \Xi(Y;\theta^c) + \partial\Xi(\theta^c)/\partial\theta(\theta^* - \theta^c)$$

- Since our model Ξ is known and θ^c is known, this expression defines an approximate model that is linear in θ^*.

  In large samples this approximation may be 'good enough'

- This type of argument should seem plausible to you. The hard part is defininig 'good enough' and proving that it is indeed 'good enough.' I'll sketch it.
- It is important to realize that application of the CAN framework starts with having a consistent estimator.

- So long as we are working with smooth models, the assumption that we have a $\theta^c \approx \theta^*$ goes a long way to explaining why the key is understanding the linear case.
- And establishing asymptotic normality (the 'AN') is pretty trivial in the linear case.

▶ **Completing the argument**

- To complete the argument, we need to figure out how to make a consistent estimator in the general case
- And then we need to show how some approximation like the one just described can form the basis for applying WLLNs and CLTs to reach our desired results.

▶ **First step: limit the discussion to the class of estimators known as 'extremum estimators' where $\hat{\theta}$ is the argmax of some objective function.**

▶ **Extremum estimators**

- In many cases, such as ML, GMM, OLS, etc., our estimator is defined as the optimum of an objective function.
- Indeed, as economists, we—like the agents in our standard models—are compulsively pulled toward decisions and methods justified as the peak of some objective function.
- We have some objective fuction, say $\Xi(Y;\theta)$ and

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \Xi(Y;\theta)$$

- We assume that Ξ is smooth so that we can use our calculus tools.
- Define $q(Y;\theta)$ and $Q(Y;\theta)$ to be the first two derivatives of the objective function.
- Our estimator will satisfy the first order condition $q(\hat{\theta};Y) = 0$.
- Take a first order approximation to q around the true parameter, θ^*,

$$q(Y;\hat{\theta}) \approx \bar{q}(Y;\theta^*) + (\hat{\theta} - \theta^*)Q(Y;\theta^*) \qquad (***)$$

  $(***)$ is the key relation. We'll return to it below

- Using $q(\hat{\theta}) = 0$, rearranging and multiplying by $\sqrt{T}$,

$$\sqrt{T}(\hat{\theta} - \theta^*) \approx -Q^{-1}(Y;\theta^*) \times \sqrt{T}q(Y;\theta^*)$$

- Dropping the arguments of the function, the RHS above is $Q^{-1} \times \sqrt{T}q$

  this replaces the identical $\bar{Q}^{-1}\bar{w}$ from the OLS case.

- Thus, follow the OLS logic: make enough assumptions that

$$Q(Y;\hat{\theta}) \rightarrow_p \mathbf{Q} \ \ (\textbf{fullrank})$$

and

$$\sqrt{T}q(\hat{\theta}) \rightarrow_d N(0, \mathbf{\Omega})$$

for some $\mathbf{\Omega}$.

- And voila:

$$\sqrt{T}(\hat{\theta} - \theta^*) \rightarrow_d N(0, \mathbf{Q^{-1}\Omega Q^{-1}})$$

- This is just as in the OLS case.

▶ So what is special about OLS?

- From the standpoint of the theory just sketched, the main thing different about OLS is that in the OLS case, the first-order approximation in (***) is exact!

  Because the objective is quadratic in OLS, the first oder approximation to the first derivative is exact.

- So long as we are not sticklers about exactness, our OLS reasoning gets us all the results we need for general extemum estimators in general smooth models.
- To emphasize: we have already done a bunch of reasoning for OLS
- As it turns out, most of that reasoning only required that (***) hold approximately rather than exactly.
- Thus, so long as all we care about is conventional (that is, first order asymptotics) we don't need to go much beyond the linear case.

  Given a $\theta^c \approx \theta^*$ a linear approximation will capture the terms that are important asymptotically.

▶ Extremum estimators a bit more fully

- The material below follows Amemiya's text and Heckman's notes, which follows Amemiya's text.

  Heckman's lecture go

  http://jenni.uchicago.edu/econ312/Slides/topic2/asymp3-extest%5F2006-04-12%5Fmms.pdf

- I'll just give the hand-wavy version as a guide to reading the more complete versions
- Suppose our objective function is $G(Y;\theta) = G_T(Y;\theta)$

- And suppose there is a nonstochastic function $\mathbf{G}(\theta)$ and

$$G(Y;\theta) \to_p \mathbf{G}(\theta)$$

for each θ

- If this is a sensible estimator, we'll need that in the limit at least, the objective function must be optimized at the true value, θ^*.

  (Without loss of generality, henceforth, I'll say maximized not optimized. Stating the problem as a maximization or minimization is merely a normalization.)

- Specifically, assume that $\mathbf{G}(\theta)$ is uniquely, globally maximized at θ^*

  θ^* being truth

▶ Aside:: Multiple maxima

- Multiple maxima is a matter of identification, which we'll discuss later.
- Thus, we are assuming global identification here.

▶ Consistency

- Under these assumptions,

$$\hat{\theta} = \operatorname*{argmax}_{\theta} G(Y;\theta)$$

is a consistent estimator for θ:

$$\hat{\theta} \to_p \theta^*$$

- Proof'ish: Suppose that our estimator has a plim (that is, it converges in probability to some constant).

  That the estimator has a plim should not trouble you given that we have assumed that $G(Y;\theta)$ has a plim for each fixed θ.

- It should be clear that if it has a plim, the plim cannot be anything but θ^*.

  Given that θ^* uniquely maximizes $\mathbf{G}$.

- That's the proof'ish. Go read the proofs to see how this works.
- This requires very few assumptions other than existence of a smooth objective that is maximized at θ^* and that various items have plims.

▶ Inference, the hand-wavy version

- Assume that the objective function, $G(Y;\theta)$, has two continuous derivatives.

  We did not need this for consistency

- In particular, define

$$\begin{aligned} q(Y;\theta_a) &= \frac{\partial G(Y;\theta_a)}{\partial \theta)} \\ Q(Y;\theta_a) &= \frac{\partial^2 G(Y;\theta_a)}{\partial \theta \partial \theta'} \end{aligned}$$

- We choose our estimator, $\hat{\theta}$, to satisify the first order condition

$$q(Y;\hat{\theta}) = 0$$

- Take a first-order approximation around θ^*:

$$q(Y;\hat{\theta}) \approx q(Y;\theta^*) + Q(Y;\theta^*)(\hat{\theta} - \theta^*)$$

- Since the LHS is zero we can re-arrange as:

$$\sqrt{T}(\hat{\theta} - \theta^*) \approx -Q^{-1}(Y;\theta^*) \times \sqrt{T} q(Y;\theta^*)$$

  I've taken the liberty of multiplying both sides by $\sqrt{T}$

- We make enough assumptions for (i) q to converge to zero in probability and (ii) for Q to converge to a full rank matrix in probability:

$$Q(Y;\theta^*) \rightarrow_p \mathbf{Q}$$

  and (iii) for $\sqrt{T}q$ to satisfy a CLT, with asymptotic variance-covariance matrix $\mathbf{\Omega}$, and voila:

$$\sqrt{T}(\hat{\theta} - \theta^*) \rightarrow_d N(0, \mathbf{Q^{-1}\Omega Q^{-1}})$$

► **Aside:: A hint to completing the proof**

- To get this all to work we need some regularity conditions and we apply a slight modification of the above sketch
- For example, we use the mean value theorem to say that $(***)$ holds exactly for some $\tilde{\theta}$ between $\hat{\theta}$ and θ^*

  And then consistency of $\hat{\theta}$ guaratees that $\tilde{\theta}$ is converging to θ^*.

► **Some examples**

- Above we made what may seem like brute assumtions that $Q()$ has a plim and $q()$ is asymptotically normal.

- Of course, about the only way we can get to results like this in practice is if $q()$ and $Q()$ are (functions of) sample means
- So we need some plausible general classes of functions to maximize that are widely adabtable to our problems.
- Maximum likelihood and GMM give us two large families here

▶ **Maximum likelihood in this framework**

- You might review Hansen's account of the MLE (especially Appendix B11) at this point

▶ **ML**

- The objective function is the log-likelihood implied by our economic model.
- The first order condition is the score, which we will generally call $q(Y;\theta)$.

▶ **iid data case**

- In the iid data case, the log-likelihood and score and Hessian are a sum of iid elements; thus, we can divide by T and have a sample mean.
- If we adjust our notation a bit, we might define $q_t()$ as the score function for each iid observation and then the MLE satisfies

$$\bar{q}(Y;\hat{\theta}_{MLE}) = 0$$

- You should know that under standard assumptions,

$$Eq_t(Y;\theta^*) = 0$$

The expectation of the score evaluted at the truth is zero.

- Thus, if the $q_t(Y;\hat{\theta}_{MLE})$ satisfy a WLLN, then the MLE should be consistent.

▶ **Hetergeneously distributed time series**

- In time series, the data are dependent and the joint likelihood need not be a sum of T identical items.
- We can sequentially condition the likelihood on the past as described earlier so that

$$\mathcal{L}(Y;\theta) = \ln(f_{\theta,1}(y_1)) + \sum_{i=2}^{T} \ln(f_{\theta,t}(y_t|y_{t-}))$$

where the t subscripts on the fs emphasizes the potential heterogeneity.

- We could always condition to write the log likelihood as a sum; just sequentially condition using any arbitrary arrangement of the obersvations.

- The resulting conditional fs need not be any easier to work with than where we started.
- Time series works because when we condition sequentially on the past, the resulting conditional distributions plausibly have a natural structure that allows us to analyze the likelihood using simple tool.
- For example, in the first order Markov case, the conditional densities, $f_{t|t-}$, are identical for all t.
- In any case, having conditioned the likelihood sequentially in this way, we have created a sum.
- And we'll make enough assumptions on good time series behavior that we can hope to repeat our WLLN- and CLT-based arguments.

► **Asymptotic normality**

- Call the Hessian of the likelihood, $Q(Y;\theta)$.
- As with the score, the Hessian can be written as a sum across individual observations.
- So let's define Q_t as the Hessian for one observation and $\bar{Q}$ as the average for all observations.
- So long as $\bar{Q}$ converges in probability to a full rank matrix and $\sqrt{T}\bar{q}$ obeys a CLT with asymptotic variance-covariance matrix $\mathbf{\Omega}$ we have

$$\sqrt{T}(\hat{\theta}_{MLE} - \theta^*) \rightarrow_d N(0, \mathbf{Q^{-1}\Omega Q^{-1}})$$

► **Aside:: Hard stuff and easy stuff conceptually**

- Making the argument just given form is pretty trivial conceptually.

  We covered the main logical bits.

- The only hard part is, e.g., figuring out various sets of extremely weak sets of sufficient conditions for the CLTs and WLLNs to hold.

  E.g., proving the theorems McFadden reports and reproduced at the end of the 'well-behaved' lecture.

- Once the theorems are proved, the asymptotic normality argument is trivial and echoes the OLS case

► **Aside:: We could go one step further**

- So far we have written:

$$\sqrt{T}(\hat{\theta}_{MLE} - \theta^*) \rightarrow_d N(0, Q^{-1}\Omega Q^{-1})$$

- But you should know that the expression for the asymptotic covariance simplifies to the inverse information matrix, call it $\mathcal{I}^{-\infty}$.

- Of course, the key to getting here is that for the MLE:

$$\Omega \equiv \lim E\bar{q}\bar{q}' = \lim -E\bar{Q} = \mathcal{I}$$

- We'll not worry about this simplification step until we get to the discussion of efficiency where this result is fundamental

  Cramer-Rao bound, etc.

▶ GMM as an extremum estimator

▶ GMM

- In GMM, we start with moment conditions implied by some model.

$$Eg(y_t;\theta) = 0$$

  where g is a vector function of ℓ conditions.

- Such moment conditions arise naturally in models of economic behavior from first order conditions for intertemporal optimization in a stochastic setting.
- For example, the consumption Euler equation under standard assumptions can be written:

$$E\beta U'(c_{t+1})/U'(c_t)(1+r_t) - 1 = 0$$

  where r_t is the return on any asset held by the person, c is consumption, β is a discount factor, and U' stands for marginal utility.

- Remembering that g is known, we have that for fixed θ, the sample analog of the moment condition is a sample mean:

$$\bar{g}(\theta) = T^{-1}\sum g(y_t;\theta)$$

- In GMM we choose θ to make $\bar{g}(\theta)$ as close to zero as possible.
- Let's take the case where the dimension of $\bar{g}$ is greater than that θ (that is, $\ell > K$)
- Thus, we won't generally be able to set all elements of the $\bar{g}$ vector to zero.
- In this case, we choose θ as,

$$\operatorname{argmin} \bar{g}(\theta)'W\bar{g}(\theta)$$

  where W is any positive definite matrix.

▶ Aside:: Identification again

- If $K = \ell$ there will generally be one θ exactly satisfying the moment conditions. We say that θ is just identified and in this case, the weight matrix is irrelevant.

- With $K > \ell$ there will be many θ satisfying the moment conditions and, thus, θ is not identified.
- We are focussing on $\ell > K$, the overidentified case.

▶ **Consistency and choice of W**

- If GMM is consistent for one positive definite W then it should be for any such matrix.
- Consistency mainly requires that θ^* optimize the objective in the limit.
- Given the quadratic form, if θ^* minimizes the objective for one positive definite matrix, it will for all.

▶ **Completing a sketch of the argument**

- Take $W = I$
- Since $\bar{g}$ is a sample mean and the population mean is zero, it should be plausible to you that

$$\sqrt{T}\bar{g}(Y;\theta^*) \rightarrow_d N(0, \mathbf{H})$$

for some $\mathbf{H}$.

▶ **Aside:: What is H?**

- Define the exact $variance - covariance$ matrix of $\sqrt{T}\bar{g}(Y;\theta)$ as

$$H_T(Y;\theta) = ET\bar{g}()\bar{g}()'$$

- Under the sort of good behavior assumptions we've been making,

$$\mathbf{H} = \lim \mathbf{H_T}() \equiv \mathbf{ET}\bar{\mathbf{g}}()\bar{\mathbf{g}}()'$$

▶ **Now (***)**

- The first order condition for minimizing $\bar{g}\bar{g}'$ is

$$\bar{D}(\hat{\theta})'\bar{g}(\hat{\theta}) = 0$$

where $\bar{D}(\theta_a) = \partial\bar{g}(\theta_a)/\partial\theta)$

And we have a bar over the D because this first derivative will also have the form of a sample mean.

- The crucial approximation, (***), in this case is:

$$\bar{g}(\hat{\theta}) \approx \bar{g}(\theta^*) + \bar{D}(\theta^*)(\hat{\theta} - \theta^*)$$

- Inserting the approximation in the FOC,

$$\bar{D}'(\theta^*)(\bar{g}(\theta^*) + D(\hat{\theta} - \theta^*)) \approx 0$$

or

$$\sqrt{T}(\hat{\theta} - \theta^*) \approx (\bar{D}'\bar{D})^{-1}\bar{D}'\bar{g}$$

- Thus, we get back to

$$\sqrt{T}(\hat{\theta} - \theta^*) \rightarrow_d N(0, \mathbf{Q^{-1}\Omega Q^{-1}})$$

where in this case we have made enough assumptions that,

$$\begin{aligned} \bar{D}\bar{D}' &\rightarrow_p \mathbf{Q} \quad (\textbf{fullrank}) \\ \mathbf{\Omega} &= \lim \bar{D}' H_T \bar{D} \end{aligned}$$

▶ **General positive definite** W

- You should be able to repeat the above argument starting with a general W in the objective.
- You'll get to:

$$\sqrt{T}(\hat{\theta} - \theta^*) \rightarrow_d N(0, \mathbf{Q^{-1}\Omega Q^{-1}})$$

where

$$\begin{aligned} \bar{D}' W \bar{D} &\rightarrow_p \mathbf{Q} \quad (\textbf{fullrank}) \\ \mathbf{\Omega} &= \lim \bar{D}' W H_T W \bar{D} \end{aligned}$$

▶ **Aside:: Efficiency**

- We are not discussing efficiency here, but you should remember that efficient GMM sets $W = H^{-1}$

  (that is, the inverse of the variance-covariance matrix of $\bar{g}$).

- And in this case, the expression for the asymptotic variance-covariance matrix simplifies.

▶ **Bottom line on CAN**

▶ **Bottom line**

- For a wide range of extremum estimators, we have a consistent estimator
- And using the fact that this estimator will be close to the true parameter asymptotically, we can rely on first order Taylor series approximations to first order conditions to reduce the general case to something that looks essentially just like the linear, OLS case.

▶ What's left

- We've been putting off that discussion of efficiency.

  And then we are largely done with the move from elementary to advanced econometrics.

- And then we can turn to the difficult issue of applying these approximations wisely in practice.

  This, I'll call the move from advanced econometrics to modern advanced econometrics.