Topic 7 : Nonlinear Models

Dr. Ani Dasgupta
Department of International Business, Mass. Maritime Academy
and
Department of Economics, Boston University

Draft lecture notes for a graduate econometrics course offered at Boston University, Spring 2018. Please do not circulate without permission.

1 Three Nonlinear Paradigms

In this topic we deal with estimation procedures for models which cannot be written as $y_i = \mathbf{x}_i{}'\boldsymbol{\beta} + \varepsilon_i$, models which will be simply referred to as 'Nonlinear Models'. However, just as a linear function is a special case of nonlinear functions, the techniques described here can be used for the classical linear model too, and with some specific assumptions made, reduce to the standard OLS technique of estimation. The procedures discussed here are asymptotic, that is they are meant to be used only when the sample size is large.

The three estimation procedures to be discussed are a) (NLS) Nonlinear Least Squares, b) (MLE) Maximum Likelihood (Maximum Likelihood Estimation) and c) GMM (Generalized Method of Moments. Each technique is apt for a certain DGP (data generating process) which we now describe. In what follows think of $\mathbf{y}$ as the variables of interest (also referred to as endogenous variables), and $\mathbf{x}$ as a vector of explanatory variables, also referred to as exogenous variables or covariates, although in some models there are no exogenous variables. These model purport to explain something about the stochasic property of $\mathbf{y}$, but typically has little to say about $\mathbf{x}$.

NLS: $E(y_i|\,\mathbf{x}_i) = \varphi(\mathbf{x}_i, \boldsymbol{\theta})$ where $\varphi()$ is some nonlinear function of $\boldsymbol{\theta}$ which is an unknown parameter vector. Equivalently, we could write $y_i = \varphi(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i$ with $E(\varepsilon_i|\mathbf{x}_i) = 0$. Typically NL models have just one endogenous variable.

Example: (Keynesian money demand function with liquidity trap) Suppose $E(M_d|\, r) = \frac{\theta_1}{r-\theta_2}$ where M_d is money demand, r is the interest rate and θ_1, θ_2 are parameters. In this model, the interest rate is restricted to be always above θ_2. A little thought should convince you that there is no way to write this relationship as a linear model.

MLE: Here, the (conditional) pdf or pmf $f(\mathbf{y_i}, \mathbf{x_i}; \boldsymbol{\theta})$ is specified. Usually $\mathbf{y}|\,\mathbf{x}$ follows a well-known distribution, the parameters of which are jointly determined by $\mathbf{x}$ and $\boldsymbol{\theta}$. Often it is either impossible or useless to write $\mathbf{y}$ in terms of a linear function of $\boldsymbol{\theta}$ and a disturbance term.

Example: (Logit Model) Let $y = 1$ if a borrower defaults on a loan and 0 otherwise. Let the probability of default be given by $\frac{e^{\boldsymbol{\theta}'\mathbf{x}}}{1+e^{\boldsymbol{\theta}'\mathbf{x}}}$ where $\mathbf{x}$ is a vector of borrower characteristics. Here y is a Bernoulli (p) variate with $p = \frac{e^{\boldsymbol{\theta}'\mathbf{x}}}{1+e^{\boldsymbol{\theta}'\mathbf{x}}}$. Again, you could not express this as a linear model. You could claim this to be an NL model since $E(y|\,\mathbf{x}) = \frac{\mathbf{e^{\boldsymbol{\theta}'x}}}{\mathbf{1+e^{\boldsymbol{\theta}'x}}}$ but just stopping at that does not capture things we know about the disturbance term (for example we know that the error variance also depends on $\mathbf{x}$) and hence leads to inefficient estimation.

GMM: In this setup we have l moment conditions

$$E[g^j(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\theta})] = 0 \quad j = 1, \dots l, \quad i = 1, \dots n$$

It is assumed that $l \geq p$ where p is the number of parameters (if $l = p$, we say that the model is exactly identified and if $l > p$, we say that it is overidentified). GMM is typically used when the data variables must satisfy moment conditions that typically emerge from the first order conditions of some maximization problem faced by economic agents.

Example: (Consumption-CAPM) Hansen and singleton (1982 Econometrica) specify a representative agent model with the agent having an intertemporal utility function $U(c_1, c_2, \cdots) = \sum_{t=0}^{\infty} \delta^t u(c_t)$, and further, for concreteness, assume that the per period utility function displays constant absolute risk aversion (CARA) by letting $u(c_t) = \frac{c_t{}^{\gamma}-1}{\gamma}$. Note that for the special case of $\gamma = 1$ this is linear in consumption and as $\gamma \to 0$ this approaches the logarithmic function. Let there be J assets, indexed by $j = 1, \dots J$, with the price of the jth asset being $p_{j,t}$, maturity being m_j and (gross) return being $r_{j,t}$. Of course $p_{j,t}$, $r_{j,t}$ are stochastic processes, but this is a rational expectation model and the agent is assumed to know the distribution of these.

In each period the agent maximizes his expected intertemporal utility subject to the budget constraint that his total wealth from his asset holding plus his labor earning must be allocated between current consumption and new asset holdings. Let $q_{j,t}$ represent the amount of the jth asset he buys in period t. In terms of algebra, this constraint can be written as

$$w_t + \sum_{j=1}^{J} r_{j,t}\, q_{j,t-m_j} = c_t + \sum_{j=1}^{J} p_{j,t}\, q_{j,t}$$

Although this appears to be a formidable dynamic programming problem, it has a simple first order condition (the so-called Euler equation):

$$p_{j,t}u'(c_t) = \delta^{m_j} E_t\left[r_{j,t+m_j}\, u'(c_{t+m_j})\right] \quad (j = 1, \dots J)$$

with the subscript t for the expectation operator denoting the idea that the expectation is being taken with respect to all that is known as of period t. The interpretetion of the Euler equation is simple: when choosing how much to invest in an asset today, the agent makes sure that marginal loss/gain of utiliy from consumption today is exactly balanced by the marginal gain/loss of utility consumption later when the asset bears return on maturity. Substituting for the CARA utility function gives

$$E_t u_t = 0 \quad (j = 1, \dots J)$$

where

$$u_t = \left[\frac{\delta^{m_j} r_{j,t+m_j}}{p_{j,t}} \left(\frac{c_{t+m_j}}{c_t}\right)^{\gamma-1} - 1\right]$$

Since $E_t u_t = 0$ also means the unconditional expectation Eu_t is also 0, thus, we have J moment conditions. In fact, one could derive many others, since if z_t is any variable known at time t, it follows that $E(u_t|z_t) = 0$ and hence $E(u_t z_t) = 0$.

Section End Questions

1. If $M = \frac{\theta_1}{r-\theta_2}$, then we could write $Mr = \theta_2 M + \theta_1$. This appears to be a linear regression model. What is wrong in estimating this to figure out what the θ's are?

2. Give an example of a stochastic optimization problem from microeconomics (it need not be dynamic). Write down the first order condition for the optimal solution, and hence exhibit a candidate GMM model.

2 Extremum Estimators

All the three estimators have the same philosophy: choose the estimator so as to maximize or minimize some criterion function. More specifically: Choose $\widehat{\theta}$ that maximizes criterion function $Q_n(\theta)$ subject to $\theta \in \Theta$(parameter space) $\subset R^p$. The criterion functions are described next:

$$NL \qquad Q_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum (y_i - \varphi(\mathbf{x}_i;\boldsymbol{\theta}))^2 \tag{1}$$

$$ML \qquad Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum \ln f(\mathbf{y}_i, \mathbf{x}_i;\boldsymbol{\theta}) \tag{2}$$

$$GMM \qquad Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{g}_n(\boldsymbol{\theta})'\widehat{\mathbf{W}}\mathbf{g}_n(\boldsymbol{\theta}) \tag{3}$$

where $\mathbf{g}_n(\theta) = \frac{1}{n}\sum \mathbf{g}(\mathbf{w}_i;\boldsymbol{\theta})$ with $\mathbf{w} = (\mathbf{y}, \mathbf{x})$ and $\mathbf{g}$ is the vector of the g^j's. Also, $\mathbf{W}$ is some suitably chosen weighting matrix.

Note: 1. that the the subscript n for the criterion function exists to remind you of the role of the sample size. 2. The first two maxes a sample average while the third doesn't, unless $\mathbf{W}$ is of a special structure. NL and ML are also called M estimators, as each is minimizing/maximizing $\frac{1}{n}\sum m(\mathbf{w}_i;\boldsymbol{\theta})$ although often in the actual optimization the n is left out. Writing the criterion as a sample average, simplifies the derivation of the properties of the M-estimators as we will see later.

Let us briefly discuss some of these properties now. Because the M-estimators are somewhat different from GMM estimator, we break up the discussion into two subsections.

2.1 M Estimators

Consistency

We state without proof the following result. We use the notation $\boldsymbol{\theta}_0$ for the true parameter vector and Q_0 denotes the expectation of m at $\boldsymbol{\theta}_0$

If a) $Q_n(\boldsymbol{\theta}) \xrightarrow[p]{} Q_0(\boldsymbol{\theta})$ for $\forall\ \boldsymbol{\theta}$ (i.e. a law of large numbers applies to criterion function)
b) $Q_0(\boldsymbol{\theta})$ is uniquely maximized at true parameter vector $\boldsymbol{\theta}_0$ (Identification Condition)
c) $Q_n(\boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$ for any dataset
d) Θ is convex
e) $Q_n(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and in data variables

then $\widehat{\boldsymbol{\theta}} \xrightarrow[p]{} \boldsymbol{\theta}_0$

Note: If concavity isn't there, we need a compact parameter space and $\sup\limits_{\boldsymbol{\theta}\in\Theta} \mid Q_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta}) \mid \xrightarrow[p]{} 0$ (a Uniform LLN applies to criterion function). In this regard note that a model may not be concave in original parameters, but a reparametization will achieve the desired result. For example : The CLM is <u>not</u> concave in (β, σ) but <u>is</u> concave in (δ, γ) where $\delta = \frac{\beta}{\sigma}$ and $\gamma = \frac{1}{\sigma}$

Identification is a serious issue in nonlinear models; so let us take a moment to discuss it in the context of NL and ML.

In the NL case,

$$Q_0(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}_0}[y_i - \varphi(x_i;\boldsymbol{\theta})]^2 \tag{4}$$

Now it is easy to prove (you might have done this in EC 507!) that the solution to the problem

$$Min\ E_{\boldsymbol{\theta}_0}[y_i - h(x_i)]$$

is $h(x_i) = E(y_i|x_i;\boldsymbol{\theta}_0)$ which of course is $\varphi(x_i;\boldsymbol{\theta}_0)$.

Hence, in this context identification requires that $\forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$

$$\varphi(x_i;\boldsymbol{\theta}) \neq \varphi(x_i;\boldsymbol{\theta}_0) \tag{5}$$

In other words it should not be the case that two different parameter vectors, creates the same conditional expectation of y given $\mathbf{x}$.

In the ML case,

$$Q_0(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0} \ln f(y_i;\mathbf{x}_i,\boldsymbol{\theta}) \tag{6}$$

We now prove alemma useful in this context.

Lemma:

$$E_{\theta_0} \ln f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0) \geq E_{\boldsymbol{\theta}_0} \ln f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}) \tag{7}$$

with equality only when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

Proof: The proof makes use of Jensen's inequalty which says that for any strictly concave function and nondegerate r.v. $f(E(x)) > E(f(x))$.

Let $X = \frac{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)}$. Then, $\ln E_{\boldsymbol{\theta}_0}[\frac{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)}] > E_{\boldsymbol{\theta}_0}[\ln(\frac{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)})]$ unless $f(\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\theta}) = f(\mathbf{y}_i,\mathbf{x}_i,\boldsymbol{\theta}_0)$ (since ln is strictly concave).

But, $\ln E_{\boldsymbol{\theta}_0}[\frac{f(y_i;\mathbf{x}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)}] = \ln 1 = 0$ ($\int \frac{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)} f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)d\mathbf{y}_i dx_i = 1$ given that $f(y_i,\mathbf{x}_i,\boldsymbol{\theta})$ is a density.)

Now the fact that $\ln[\frac{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta})}{f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)}] = \ln f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}) - \ln f(\mathbf{y}_i;\mathbf{x}_i,\boldsymbol{\theta}_0)$ gives the result. ♣

This shows that the identification condition here is

$$\text{for } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \qquad f(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\theta}) \neq f(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\theta}_0). \tag{8}$$

As examples of unidentified NLS model, first consider $y_i = \beta_1 e^{\beta_2 + \beta_3 x} + \varepsilon_i$. Since if $\beta_3 = \widetilde{\beta}_3$ & $\widetilde{\beta}_1 e^{\widetilde{\beta}_2} = \beta_1 e^{\beta_2}$ we have a violation of identification condition. Another example is $y_i = \beta_1 + \beta_2 x^{\beta_3} + \varepsilon_i$. If β_2 or β_3 is allowed to be 0, we're in trouble!

Asymptotic Normality

Recall that the criterion function is given by

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum m(\mathbf{w}_i, \boldsymbol{\theta})$$

We also define the following objects which are simply the gradient vector and the Hessian matrices of the criterion function.

$$\mathbf{S}(\mathbf{w}_i, \boldsymbol{\theta}) = \frac{\partial m(\mathbf{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{9}$$

$$\begin{aligned}\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}) &= \frac{\partial S(\mathbf{w}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\ &= \frac{\partial^2 m(\mathbf{w}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\end{aligned} \tag{10}$$

(It's symmetric matrix, thus order won't matter)

By definition of the extremum estimator $\widehat{\boldsymbol{\theta}}$,

$$\begin{aligned}\mathbf{0} &= \frac{\partial Q_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \\ &= \frac{1}{n}\sum \mathbf{S}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})\end{aligned} \tag{11}$$

Using the Mean Value Theorem from multivariable calculus: [1]

$$\begin{aligned}\frac{\partial Q_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} &= \frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 Q_n(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= \frac{1}{n}\sum \mathbf{S}(\mathbf{w}_i; \boldsymbol{\theta}_0) + \frac{1}{n}\sum \mathbf{H}(\mathbf{w}_i; \widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= 0 \ (\because \widehat{\boldsymbol{\theta}} \text{ maxes } Q_n)\end{aligned}$$

$$\therefore \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -[\frac{1}{n}\sum \mathbf{H}(\mathbf{w}_i; \widetilde{\boldsymbol{\theta}})]^{-1}\frac{1}{\sqrt{n}}(\sum \mathbf{S}(\mathbf{w}_i, \boldsymbol{\theta}_0)) \tag{12}$$

(The Hessian will be negative definite in the neighborhood of θ_0; otherwise $\boldsymbol{\theta}_0$ will not be the unique maximizer of $Q_0(\boldsymbol{\theta})$)

Under 'suitable technical conditions' to guarantee a WLLN type of result[2]

[1] If $h : R^p - R^q$, $\underset{q\times 1}{h(X)} = \underset{q\times 1}{h(X_o)} + \underset{q\times p}{(\frac{\partial h}{\partial X'})}|_{X_0}\underset{p\times 1}{(X - X_0)}$

[2] $E[\sup_{\boldsymbol{\theta}}\|\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta})\|] < \infty$ will do for instance.

$$\left[\frac{1}{n}\sum \mathbf{H}(\mathbf{w}_i;\widetilde{\boldsymbol{\theta}})\right] \underset{p}{\longrightarrow} E\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right] \equiv \boldsymbol{\Psi} \tag{13}$$

Under further 'suitable technical conditions' to guarantee a CLT like result[3]

$$\frac{1}{\sqrt{n}}\sum \mathbf{S}(\mathbf{w}_i;\boldsymbol{\theta}_0) \underset{d}{\longrightarrow} N(0,\boldsymbol{\Sigma}) \;\; \text{where } \boldsymbol{\Sigma} = Var(\mathbf{S}(\mathbf{w}_i;\boldsymbol{\theta}_0)) \tag{14}$$

Then,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{d}{\longrightarrow} N(\mathbf{0}, \boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}) \tag{15}$$

We next turn to the issue of figuring out what the asymptotic variance matrices are and how to consistently evaluate them.

The NL Case: What is $\boldsymbol{\Psi}$ and what is $\boldsymbol{\Sigma}$ in this context?

$$Q_n = -\frac{1}{n}\sum \underbrace{(y_i - \varphi(\mathbf{x}_i;\boldsymbol{\theta}))}_{m(\mathbf{w}_i;\boldsymbol{\theta})}{}^2$$

$$\mathbf{S}(\mathbf{w}_i;\boldsymbol{\theta}) = 2\begin{pmatrix} (y_i - \varphi(\mathbf{x}_i;\boldsymbol{\theta}))\frac{\partial\varphi}{\partial\theta_1} \\ \vdots \\ (y_i - \varphi(\mathbf{x}_i;\boldsymbol{\theta}))\frac{\partial\varphi}{\partial\theta_p} \end{pmatrix} \tag{16}$$

$$\mathbf{S}(\mathbf{w}_i;\boldsymbol{\theta}_0) = 2\varepsilon_i\mathbf{X}_i \quad \text{where } \mathbf{X}_i = \begin{pmatrix} \frac{\partial\varphi}{\partial\theta_1} \\ \vdots \\ \frac{\partial\varphi}{\partial\theta_p} \end{pmatrix}_{\boldsymbol{\theta}_0}$$

Hence, $\boldsymbol{\Sigma} = Var(2\,\varepsilon_i\mathbf{X}_i)$

If we assume conditional homoskedsticity ($E(\varepsilon_i^2|\mathbf{x}_i) = \sigma^2$), this can be rewritten as

$$\begin{aligned} \boldsymbol{\Sigma} &= 4E(\varepsilon_i^2\mathbf{X}_i\mathbf{X}_i') - 4E(\varepsilon_i\mathbf{X}_i)(E(\varepsilon_i\mathbf{X}_i))' \\ &= 4\sigma^2E(\mathbf{X}_i\mathbf{X}_i') \;\; (\because E(\varepsilon_i\mathbf{X}_i) = \mathbf{0} \text{ as } E(\varepsilon_i|\mathbf{X}_i) = \mathbf{0}) \end{aligned} \tag{17}$$

Now, consider $\mathbf{H}$ and $\boldsymbol{\Psi}$. The $(j-k)$th element of $\mathbf{H}$ is

$$\mathbf{H}_{jk}(\mathbf{w}_i;\boldsymbol{\theta}_0) = 2\varepsilon_i\frac{\partial^2\varphi}{\partial\theta_j\partial\theta_k}|_{\boldsymbol{\theta}_0} - 2\frac{\partial\varphi}{\partial\theta_j}|_{\boldsymbol{\theta}_0}\frac{\partial\varphi}{\partial\theta_k}|_{\boldsymbol{\theta}_0}$$

[3] Finite variance of the random vector in question will do.

Upon taking expectation

$$\mathbf{\Psi}_{jk} = -2E\left[\left(\frac{\partial\varphi}{\partial\theta_j}\right)_{\boldsymbol{\theta}_0}\left(\frac{\partial\varphi}{\partial\theta_k}\right)_{\theta_0}\right]$$

Hence,

$$\mathbf{\Psi} = -2E(\mathbf{X}_i\mathbf{X}_i') \tag{18}$$

Combining (15), (17) and (18), we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{d}{\longrightarrow} N(0, \sigma^2\left[E(\mathbf{X}_i\mathbf{X}_i')\right]^{-1}) \tag{19}$$

NOTE: If φ is actually linear, ie. $\varphi(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}_1 x_1 + \cdots + \theta_p x_p$, then $\mathbf{X}_i$ is actually $\mathbf{x}_i$ and we have our standard result that under conditional homoskedasticity,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{d}{\longrightarrow} N(0, \sigma^2{\mathbf{\Sigma}_x}^{-1}) \tag{20}$$

For implementing (asymptotic) tests as well creating confidence intervals, both σ^2 and $E(\mathbf{X}_i\mathbf{X}_i')$ can be (consistently) estimated by corresponding sample averages as usual.

The ML Case: What is $\mathbf{\Psi}$ and what is $\mathbf{\Sigma}$ here?

Noting that $m(\mathbf{w}_i, \boldsymbol{\theta}) = lnf(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})$, we proceed exactly as before, but let's first state some preliminary results.

$$E_{\boldsymbol{\theta}_0}\left[\frac{f(\mathbf{y},\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{y},\mathbf{x};\boldsymbol{\theta}_0)}\right] = 1 \tag{21}$$

$$E_{\boldsymbol{\theta}_0}\left[\frac{\frac{\partial f}{\partial\theta_j}(\mathbf{y},\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{y},\mathbf{x};\boldsymbol{\theta}_0)}\right] = 0 \tag{22}$$

$$E_{\boldsymbol{\theta}_0}\left[\frac{\frac{\partial^2 f(\mathbf{y},\mathbf{x};\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}}{f(\mathbf{y},\mathbf{x};\boldsymbol{\theta}_0)}\right] = 0 \tag{23}$$

Equation (21) follows from the fact that $f(y, \mathbf{x}; \theta)$ is a density for every $\boldsymbol{\theta}$, (22) follows from differentiating (21) with respect to θ_j and (23) in turn follows from differentiating (22) with respect to θ_k.

In the ML context $\mathbf{S}(\mathbf{w}_i, \boldsymbol{\theta}) = \frac{\partial \ln f}{\partial \boldsymbol{\theta}}$ and it is called the score vector (the j'th entry of which is $\frac{\partial f/\partial\theta_j}{f}$).

Note that because of (22),

$$E_{\boldsymbol{\theta}_0}\mathbf{S}(\mathbf{w}_i,\boldsymbol{\theta}_0)=0 \tag{24}$$

Now lets evaluate the Hessian which is

$$\mathbf{H}=\frac{\partial^2\ln f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$$
$$\mathbf{H}_{jk}=\frac{\frac{\partial^2 f(y,\mathbf{x};\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}}{f(y,\mathbf{x};\boldsymbol{\theta})}-\frac{\frac{\partial f(y,\mathbf{x};\boldsymbol{\theta})}{\partial\theta_j}\times\frac{\partial f(y,\mathbf{x};\boldsymbol{\theta})}{\partial\theta_k}}{f(y,\mathbf{x};\boldsymbol{\theta})^2}$$

Define

$$\mathbf{I}_i(\boldsymbol{\theta}_0):=-E_{\boldsymbol{\theta}_0}\left[\mathbf{H}(\mathbf{w}_i,\boldsymbol{\theta}_0)\right]=E_{\boldsymbol{\theta}_0}\left[\mathbf{S}(\mathbf{w}_i,\boldsymbol{\theta}_0)\mathbf{S}(\mathbf{w}_i,\boldsymbol{\theta}_0)'\right]=Var_{\boldsymbol{\theta}_0}\left[\mathbf{S}(\mathbf{w}_i,\boldsymbol{\theta}_0)\right] \tag{25}$$

where we are making use of (23) and (24).

Going back to equation (15), we see that in the ML case

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\underset{d}{\longrightarrow}N(0,\mathbf{I}_i^{-1}) \tag{26}$$

since $\boldsymbol{\Sigma}=Var(\mathbf{S}(\boldsymbol{\theta}_0))=\mathbf{I}_i$ and $\boldsymbol{\Psi}=-\mathbf{I}_i$.

So, there is a simplification of the asymptotic variance expression just like in the NL case.

There are 3 Ways of Estimating $\mathbf{I}_i$; let me describe them next.

1. Write down the expression for $E\frac{\partial^2\ln f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$ (this will typically involve tons of theoretical calculations), and then plug in estimated parameters.

$$\widehat{\mathbf{I}}_i=\left[-E_{\boldsymbol{\theta}}\frac{\partial^2\ln f(\mathbf{w}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\widehat{\boldsymbol{\theta}}} \tag{27}$$

2. Instead of taking expectation of $\frac{\partial^2\ln f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$, just use sample averages.

$$\widehat{\mathbf{I}}_i=\left[-\frac{1}{n}\sum\frac{\partial^2\ln f(\mathbf{w}_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]_{\widehat{\boldsymbol{\theta}}} \tag{28}$$

3. This estimator (called the BHHH estimator after its inventors) or the ‘Outer estimator’) avoids even taking second partials.

$$\widehat{\mathbf{I}}_i=\frac{1}{n}\sum\mathbf{S}(\mathbf{w}_i,\widehat{\boldsymbol{\theta}})\mathbf{S}(\mathbf{w}_i,\widehat{\boldsymbol{\theta}})' \tag{29}$$

This makes use of equation (25).

ML for the Classical Linear Model w/ Normal Disturbances

Here we have: $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\sigma}^2 I), f(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}$. Hence,

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \\ \ln L &= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \frac{\partial \ln L}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = 0 \\ \frac{\partial \ln L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \end{aligned}$$

Solving these equations gives

$$\begin{aligned} \mathbf{b}_{MLE} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \sigma^2_{MLE} &= \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n} \end{aligned} \tag{30}$$

Therefore, $\mathbf{b}_{MLE}$ coincides with the OLS estimator for $\boldsymbol{\beta}$, but the estimate of σ^2 is different in this setup.

Now, let us calculate

$$\begin{aligned} \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} \\ \frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \sigma^2} &= \frac{1}{2\sigma^4}\left[-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right] \\ \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

Define $\boldsymbol{\mathcal{H}}(\mathbf{w}, \boldsymbol{\theta})$ to be the Hessian of the (entire) loglikelihood function, and $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ to be the negative of its expectation (it should be clear that $\boldsymbol{\mathcal{H}}(\mathbf{w}, \boldsymbol{\theta}) = \sum \mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta})$ and $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \sum \mathbf{I}_i(\boldsymbol{\theta}) = n\mathbf{I}_i$.)

Thus,

$$\begin{aligned} E\boldsymbol{\mathcal{H}} &= \begin{bmatrix} -\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix} \\ \boldsymbol{\mathcal{I}} &= \begin{bmatrix} \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \end{aligned}$$

$$\boldsymbol{\mathcal{I}}^{-1} = \left[\begin{array}{cc} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & \frac{2}{n}\sigma^4 \end{array} \right] \tag{31}$$

The celebrated Cramer-Rao Inequality states that the variance-covariance matrix of any unbiased estimate of θ exceeds $\boldsymbol{\mathcal{I}}^{-1}$ by a positive semidefinite matrix. Thus $\boldsymbol{\mathcal{I}}^{-1}$ is in a sense the 'lower bound' of the variances of all unbiased estimators of $\boldsymbol{\theta}$. What we have just shown is that the variance of the OLS estimate does reach its (corresponding) CR lower bound!

2.2 Generalized Method of Moment(GMM)

Let, for each observation, it be true that

$$Eg^j(\mathbf{w}_i, \boldsymbol{\theta}) = 0 \qquad j = 1, ..., l$$

The criterion function for GMM estimator is

$$Q_n(\theta) = -\frac{1}{2}\mathbf{g}_n(\boldsymbol{\theta})'\widehat{\mathbf{W}}\mathbf{g}_n(\boldsymbol{\theta}) \tag{32}$$

where $\mathbf{g}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum \mathbf{g}(\mathbf{w}_i, \theta)$, $\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) = \begin{pmatrix} g^1(\mathbf{w}_i, \boldsymbol{\theta}) \\ \vdots \\ g^l(\mathbf{w}_i, \boldsymbol{\theta}) \end{pmatrix}$ and $\widehat{\mathbf{W}}$ is some symmetric, p.d. matrix (possibly data generated) that converges to a symmetric p.d. matrix.

We now show that two of our familiar estimators can be cast in the light of a GMM estimator.

1. $\mathbf{b}_{OLS}$ is a GMM estimator:

We have $E\mathbf{x}_i(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta}) = 0$ (since $\mathbf{x}_i$ is a k-vector, thus there are k moment conditions)

Hence, $\mathbf{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum(\mathbf{x}_i y_i - \mathbf{x}_i\mathbf{x}_i'\boldsymbol{\beta}) = \frac{1}{n}(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta})$

We let $\widehat{\mathbf{W}} = \mathbf{I}$

Hence, our criterion function (ignoring the multiplicative constant $\frac{1}{n}$) is $-(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta})$. Clearly this is maximized if $(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta})$ can be set to $\mathbf{0}$, which is exactly what the OLS estimator does.

2. $\mathbf{b}_{2SLS}$ is a GMM estimator :

In this setup, we have $E\mathbf{w}_i(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta}) = 0$ Note that here $\mathbf{w}_i$ is the l-dimensional vector of instruments in keeping with the notation from Topic 6 and not $(y_i, \mathbf{x}_i)$. Thus there are l conditions.

Hence, $\mathbf{g}_n(\boldsymbol{\beta}) = \frac{1}{n}(\mathbf{W}'\mathbf{y} - \mathbf{W}'\mathbf{X}\boldsymbol{\beta})$

We choose $\widehat{\mathbf{W}} = (\mathbf{W}'\mathbf{W})^{-1}$

Hence, our criterion function is (ignoring multiplicative constants)

$$\begin{aligned}
&- (\mathbf{W}'\mathbf{y} - \mathbf{W}'\mathbf{X}\boldsymbol{\beta})'(\mathbf{W}'\mathbf{W})^{-1}(\mathbf{W}'\mathbf{y} - \mathbf{W}'\mathbf{X}\boldsymbol{\beta})\\
&= -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\\
&= -(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{P}_{\mathbf{W}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\\
&= -(\mathbf{P}_{\mathbf{W}}\mathbf{y} - \mathbf{P}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta})'(\mathbf{P}_{\mathbf{W}}\mathbf{y} - \mathbf{P}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta})
\end{aligned}$$

Now, it is a least squares problem where we are working with $\mathbf{P_w y}$ and $\mathbf{P_w X}$ instead of $\mathbf{y}, \mathbf{X}$ and hence, clearly, the argmax is $(\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{X})^{-\mathbf{1}}(\mathbf{X}'\mathbf{P}_{\mathbf{W}}\mathbf{y})$, the 2SLS estimator!

Consistency

We present the following necessary condition without proof. We revert back to the notation where $\mathbf{w}_i$ represents the data vector for the i-th observation.

If the parameter space is compact, $\mathbf{g}$ is continuous in $\mathbf{w}$ and $\boldsymbol{\theta}$.

$$\begin{aligned}
E\mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) \neq 0 \quad for\ \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \qquad &\text{(Identifiability condition)}\\
E\left[\sup_{\boldsymbol{\theta}\in\Theta} \parallel \mathbf{g}(\mathbf{w}_i, \boldsymbol{\theta}) \parallel\right] < \infty \qquad &\text{(Dominance condition)}
\end{aligned}$$

then $\widehat{\boldsymbol{\theta}} \underset{p}{\longrightarrow} \boldsymbol{\theta}_0$.

Compactness can be guaranteed by bounding the parameter space, and the moment functions are usually continuous in data and parameters. The identifiability condition is also most likely to be satisfied in practice. However, the last condition must be taken on faith...

Asymptotic Normality

We begin by stating a result on matrix differentiation.

Let $f(\boldsymbol{\theta}) = \mathbf{X}(\boldsymbol{\theta})'\mathbf{A}\mathbf{X}(\boldsymbol{\theta})$

where $\boldsymbol{\theta}$ is $p \times 1$, $\mathbf{X} : \mathbb{R}^p \longrightarrow \mathbb{R}^l$, $\mathbf{A}$ is $l \times l$ and $f : \mathbb{R}^p \longrightarrow \mathbb{R}^l$

then, $\frac{\partial f}{\partial \boldsymbol{\theta}} = 2\left(\frac{\partial \mathbf{X}}{\partial \boldsymbol{\theta}'}\right)' \mathbf{A}\mathbf{X}(\boldsymbol{\theta})$

where $\left(\frac{\partial \mathbf{X}}{\partial \boldsymbol{\theta}'}\right)$ is the $l \times p$ matrix of $\frac{\partial \mathbf{X}_i}{\partial \boldsymbol{\theta}_j}'s$.

You should be able to verify this by using Chain Rule.

Now, since $\frac{\partial Q_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = 0$, from (32) we see that

$$-\mathbf{G}_n(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{W}}\mathbf{g}_n(\widehat{\boldsymbol{\theta}}) = 0 \tag{33}$$

where $\mathbf{G}_n(\boldsymbol{\theta})$ is the Jacobian of the moment conditions w.r.t $\boldsymbol{\theta}$, i.e.

$$\mathbf{G}_n(\boldsymbol{\theta}) = \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \quad \text{which is } l \times p \tag{34}$$

Expanding $\mathbf{g}_n(\widehat{\boldsymbol{\theta}})$ about $\mathbf{g}_n(\boldsymbol{\theta}_0)$, we have

$$\mathbf{g}_n(\widehat{\boldsymbol{\theta}}) = \mathbf{g}_n(\boldsymbol{\theta}_0) + \mathbf{G}_n(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{35}$$

where $\widetilde{\boldsymbol{\theta}}$ lies on the line segment joining $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.

Multiplying both sides of (35) w/ $G_n(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{W}}$ and making use of (33), we have

$$\mathbf{G}_n(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{W}}\mathbf{g}_n(\widehat{\boldsymbol{\theta}}) = 0 = \mathbf{G}_n(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{W}}\mathbf{g}_n(\boldsymbol{\theta}_0) + \mathbf{G}_n(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{W}}\mathbf{G}_n(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{36}$$

Hence

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\left[\mathbf{G}_n(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{W}}\mathbf{G}_n(\widetilde{\boldsymbol{\theta}})\right]^{-1} \mathbf{G}_n(\widehat{\boldsymbol{\theta}})'\widehat{\mathbf{W}}\sqrt{n}\left(\frac{1}{n}\sum g(\mathbf{w}_i, \boldsymbol{\theta})\right)$$

Under suitable technical conditions[4]

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{d}{\longrightarrow} N(0, (\mathbf{G}'\widehat{\mathbf{W}}\mathbf{G})^{-1}(\mathbf{G}'\widehat{\mathbf{W}}\mathbf{S}\widehat{\mathbf{W}}\mathbf{G})(\mathbf{G}'\widehat{\mathbf{W}}\mathbf{G})^{-1})$$

where $\mathbf{G} = E\left[\mathbf{G}_n(\boldsymbol{\theta}_0)\right]$ and $\mathbf{S} = Var(g(\mathbf{w}_i, \boldsymbol{\theta}))$.

As usual, $\mathbf{G}$ and $\mathbf{S}$ can be estimated easily using sample counterparts.

The choice of $\widehat{\mathbf{W}}$ and 2-step GMM :

So far, we have been silent about the choice of $\widehat{\mathbf{W}}$. It turns out that a suitable choice of $\widehat{\mathbf{W}}$, we can create the most efficient GMM estimator among the class of all estimators. That choice is $\widehat{\mathbf{W}} = \mathbf{S}^{-1}$(Can you prove this?)

[4] These include, for instance, assumptions needed for the consistency of $\widehat{\theta}$, that $\mathbf{G}_n(\boldsymbol{\theta}_0)$ is of full column rank, $E\left[\sup_{\boldsymbol{\theta}\in\Theta} \parallel \frac{\partial g}{\partial \boldsymbol{\theta}'} \parallel\right] < \infty$ and $g(\mathbf{w}_i; \boldsymbol{\theta})$ has finite variance.

However, this creates a conundrum: unless we already have an estimator, we can't estimate $\mathbf{S}$ or $\mathbf{S}^{-1}$ and at the same time we need $\mathbf{S}$ to evaluate the criterion function and optimize it.

The conundrum is solved by choosing a 2 step procedure as follows.

Step 1: Choose any $\widehat{\mathbf{W}}$ (often the choice is simply $\mathbf{I}$) which is symmetric, p.d. and converges in probability to a symmetric p.d. matrix. Evaluate $\widehat{\boldsymbol{\theta}}$ and hence $\mathbf{g}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})'s$ and estimate $\mathbf{S}$ by calculating the sample variances of the $\mathbf{g}'s$. Let this be $\widehat{\mathbf{S}}$.

Step 2: Now use $\widehat{\mathbf{S}}^{-1}$ as $\widehat{\mathbf{W}}$ and optimize the criterion function again to obtain the efficient GMM estimator.

As a consequence of this choice, we obtain

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{d}{\longrightarrow} N(\mathbf{0}, (\mathbf{G}'\mathbf{S}^{-1}\mathbf{G})^{-1})$$

Section End Questions

1. Give a pictorial intuition behind Jensen's inequality.
2. For the Keynesian money demand example, explicitly calculate the asymptotic variance of the NLS estimator.
3. For the logit model, explicitly calculate the formulas of all three estimators of the asymptotic variance.

3 Tests

In the setups we have just discussed, we can create (asymptotic) tests for a general (possibly nonlinear) hypothesis of the form $H_0 : \mathbf{a}(\boldsymbol{\theta}_0) = \mathbf{0}$ against $H_0 : \mathbf{a}(\boldsymbol{\theta}_0) \neq \mathbf{0}$ where $\mathbf{a}$ is an $r \times 1$ function. We use $\mathbf{A}(\boldsymbol{\theta})$ to denote $\frac{\partial \mathbf{a}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$. We will assume $\mathbf{A}_0 = \frac{\partial \mathbf{a}(\theta_0)}{\partial \boldsymbol{\theta}'}$ is of full row rank. In what follows $\widehat{\boldsymbol{\theta}}_U$ will denote the unrestricted estimator and $\widehat{\boldsymbol{\theta}}_R$ will denote the restricted estimator.

There are 3 basic tests: the Wald test, the Lagrange Multiplier test and the Likelihood Ratio test (the last one is a misnomer when applied in GMM context)

- The intuition behind the Wald test is that if H_o was true then $\mathbf{a}(\widehat{\boldsymbol{\theta}}_U)$ should be 'pretty close to' 0.
- The intuition behind the LM test is that if H_o was true, then the average score vector $\left(\frac{\partial Q_n}{\partial \boldsymbol{\theta}}\right)$ evaluated at $\widehat{\boldsymbol{\theta}}_R$ should be 'pretty close to' 0.
- The intuition behind the LR test is that if H_o was true, the optimized criterion function for the restricted model should be 'pretty close to' the optimized criterion for that of the unrestricted model.

  (See Green Fig. 16.2 for a pictorial intuition)

I now present the test statistics under the null in each case. While the Wald statistic is a straightforward application of the Delta Method, for the arguments behind the other two see Hayashi Ch.7). For the GMM case the understanding is that we are referring to the efficient GMM estimator (i.e. $\widehat{\mathbf{W}}$ has been chosen optimally).

$$WALD : na(\widehat{\boldsymbol{\theta}}_U)\left[A(\widehat{\boldsymbol{\theta}}_u)\widehat{\boldsymbol{\Gamma}}_u A(\widehat{\boldsymbol{\theta}}_u)\right]^{-1}\mathbf{a}(\widehat{\boldsymbol{\theta}}_u)$$

$$LM : n\left(\frac{\partial Q_n(\widehat{\boldsymbol{\theta}}_R)}{\partial \boldsymbol{\theta}}\right)'\widehat{\boldsymbol{\Gamma}}_R\left(\frac{\partial Q_n(\widehat{\boldsymbol{\theta}}_R)}{\partial \theta}\right)$$

$$LR : 2n\left[Q_n(\widehat{\boldsymbol{\theta}}_U) - Q_n(\widehat{\boldsymbol{\theta}}_R)\right]$$

where $\widehat{\boldsymbol{\Gamma}}$ is an estimator for $Avar(\widehat{\boldsymbol{\theta}})$. Note that

$$\begin{array}{rl}\text{For NL} & \widehat{\boldsymbol{\Gamma}} = \sigma^2 E(\mathbf{X}_i\mathbf{X}_i')^{-1} \text{ (under conditional homoskedasticity)} \\ \text{For ML} & \widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{I}}_i^{-1} \\ \text{For (efficient) GMM} & \widehat{\boldsymbol{\Gamma}} = (\mathbf{G}'\widehat{\mathbf{S}}^{-1}\mathbf{G})^{-1}\end{array}$$

(whether it is $\widehat{\boldsymbol{\Gamma}}_U$ or $\widehat{\boldsymbol{\Gamma}}_R$ depends on whether we are using the unrestricted or restricted model.)[5]

[5] Incidentally, the LM statistic above reduces to the uncentered R^2 obtained when 1 is regressed on the scores in the ML model.

Note that for the Wald test, it suffices to evaluate just the unrestricted model, for the LM test just the restricted model and for the LR test both model need to be estimated. Finally, under the null each of these statistics has an asymptotic chi-squared distribition with r degrees of freedom.

A Simple Worked Out Example in the ML Context

Let us go back to the ML Estimation of a Poisson Model where

$f(y_i;\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}$

and

$(y_1, ..., y_n) = (5, 0, 1, 1, 0, 3, 2, 3, 4, 1)$

Suppose we wish to test $H_0 : \theta = 1.8$

Now $\ln f(y_i, \theta) = -\theta + y_i \ln\theta - \ln y_i!$

$\sum \ln f(y_i, \theta) = -n\theta + \ln\theta \sum y_i - \sum \ln y_i!$

The FOC is $\quad -n + \frac{\sum y_i}{\theta} = 0$

$\therefore \widehat{\theta}_U = \frac{\sum y_i}{n} = \overline{y}$

Now, $\frac{\partial^2 \ln f}{\partial\theta^2} = -\frac{y_i}{\theta^2}$

$\therefore E\left[\frac{\partial^2 \ln f}{\partial\theta^2}\right] = -\frac{1}{\theta}$ $(\because Ey_i = \theta)$

Hence, $I_i = \frac{1}{\theta}$ [We are using the first method here.]

In our example, $a(\theta) = \theta - 1.8$

Hence, $A(\theta)$ is simply 1.

Also, $\frac{\partial Q_n(\theta)}{\partial\theta} = \frac{1}{n}\frac{\partial}{\partial\theta}\sum \ln f = \frac{1}{n}\sum(-1 + \frac{\theta}{y_i}) = -1 + \frac{\overline{y}}{\theta}$

We are now ready to create the test statistics.

Wald Statistic:

$$= \underset{n}{\underset{\downarrow}{10}}(\underset{a(\widehat{\theta}_u)}{\underset{\downarrow}{2 - 1.8}})'[\ \underset{A(\widehat{\theta}_u)}{\underset{\downarrow}{1}} \cdot \underset{\widehat{I}_i(\widehat{\theta}_u)^{-1}}{\underset{\downarrow}{2}} \cdot 1]^{-1}(2 - 1.8)$$

$= 0.2$

LM Statistic:

$$= \underset{n}{\underset{\downarrow}{10}}(\underset{\frac{\partial Q_n(\widehat{\theta}_R)}{\partial \theta}}{\underset{\downarrow}{-1 + \frac{2}{1.8}}})'(\ \underset{\widehat{I}_i(\widehat{\theta}_R)^{-1}}{\underset{\downarrow}{1.8}}\ )(-1 + \frac{2}{1.8})$$

$= 0.22$

LR Statistic:

$= 2 \times 10 \times [-1.83793 + 1.84865]$

$= 0.21$

In this problem $\boldsymbol{\theta}$ was one-dimensional, so the calculations were relatively straightforward. In an upcoming problem set I will ask you to do testing when $\boldsymbol{\theta}$ has more than one parameter. However, the principle is identical. Also, please see the handout called “ML Estimation in STATA”’ to know how to utilize STATA to do testing in the ML context using STATA’s built-in constructs.