NONLINEAR ESTIMATION: M-ESTIMATION

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction and Examples
2. Consistency of M-Estimators
3. Asymptotic Distribution
4. Estimating the Asymptotic Variance
5. Large-Sample Inference
6. Two-Step M-Estimators
7. Bootstrapping

1. INTRODUCTION AND EXAMPLES

- Up until now, all estimators we have studied can be written as “closed form” functions of the data. That is, given the observed data, we have a mathematical rule for obtaining the estimate. For example, the OLS estimator is

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i' y_i \right).$$

- Such estimators do not cover all cases of interest, particularly when we turn to nonlinear models.

- Even if the underlying model is linear, special asymptotic methods are sometimes needed for certain estimators. Suppose that

$$Med(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} = \beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K$$

is the conditional median of y given $\mathbf{x}$. Without additional assumptions, OLS does not consistently estimate the β_j. But *least absolute deviations* (LAD) does. The LAD estimator solves

$$\min_{\mathbf{b} \in \mathbb{R}^K} \sum_{i=1}^{N} |y_i - \mathbf{x}_i \mathbf{b}|,$$

that is, it minimizes the sum of absolute deviations (or residuals).

- For large-sample analysis, a key point is that the LAD estimate cannot generally be written in closed form.
- Suppose that a solution to the problem does exist; call it $\hat{\boldsymbol{\beta}}_{LAD}$. Then we know

$$\hat{\boldsymbol{\beta}}_{LAD} = \mathbf{g}(\mathbf{x}_1, y_1, \mathbf{x}_2, y_2, \ldots, \mathbf{x}_N, y_N)$$

for some function $\mathbf{g}(\bullet)$. But we do not know $\mathbf{g}(\bullet)$.

- Question: If we do not know $\mathbf{g}(\bullet)$, how do we study the large-sample (asymptotic) properties of $\hat{\boldsymbol{\beta}}_{LAD}$?
- Answer: Indirectly, through the properties of the objective function.

- In particular, for each (dummy argument) **b**,

$$N^{-1}\sum_{i=1}^{N}|y_i - \mathbf{x}_i\mathbf{b}|$$

is an average of independent, identically distributed random variables, $q_i(\mathbf{b}) \equiv |y_i - \mathbf{x}_i\mathbf{b}|$, and so we can apply the law of large numbers if $E[q_i(\mathbf{b})] < \infty$.

- As another example, suppose for $y \geq 0$ we specify an exponential conditional mean model:

$$E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta}) = \exp(\beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K).$$

- Without further assumptions, we cannot “linearize” the model by using $\log(y)$ as the dependent variable. (In fact, $\log(y)$ may not even be well defined.)

- Instead, we can directly estimate $\boldsymbol{\beta}$ by *nonlinear least squares* (NLS):

$$\min_{\mathbf{b}\in\mathbb{R}^K} \sum_{i=1}^{N} [y_i - \exp(\mathbf{x}_i\mathbf{b})]^2.$$

- As in the case of LAD, we cannot present the solution in closed form. But the estimator minimizes a function that is an average of i.i.d. random functions of **b**.
- For our purposes, “nonlinear” means any situation where an estimator cannot be obtained in closed form. This requires a new set of tools for asymptotic analysis.

2. CONSISTENCY OF M-ESTIMATORS

- We first cover a class of estimation problems estimators known as *M-estimation*. (The “M” refers, for us, to “minimization.” Originally, M-estimators we defined as maximization problems.)
- We will carry along the example of nonlinear least squares for a general regression function. Because we will require a separate notation for the value of the parameters describing the population, and the set of candidates for those values, we introduce a new convention.

- Consider a linear regression model where we know the population values, say

$$E(y|x,z) = 3.26 + 0.75\, x - 1.84\, z.$$

This population regression is a particular version of the *model*

$$m(x,z) = \theta_1 + \theta_2 x + \theta_3 z$$

for $E(y|x,z)$, where each θ_j can range across all real numbers.

- It is helpful to let $(\theta_1, \theta_2, \theta_3)$ denote a generic candidate for the actual population parameters. The actual population parameters are denoted θ_{01}, θ_{02}, and θ_{03}. That is, $\theta_{01} = 3.26$, $\theta_{02} = 0.75$, and $\theta_{03} = -1.84$.

- In practice, of course, we do not know the vector of values, $\boldsymbol{\theta}_o$, actually describing the population. But it is these constants that we hope to estimate. In order to state assumptions that allow us to do so in general nonlinear contexts, we need to distinguish between $\boldsymbol{\theta}_o$ and a generic vector, $\boldsymbol{\theta}$.
- For NLS, we specify a model for the conditional mean, $E(y|\mathbf{x})$, where y is a scalar response and $\mathbf{x}$ is a vector. We focus on *parametric* models, which means the function is known up to the unknown parameters. Let $m(\mathbf{x}, \boldsymbol{\theta})$ represent this function for all $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^P$, where P is a positive integer.

- So $\boldsymbol{\theta}$ is a $P \times 1$ vector. The *parameter space* $\boldsymbol{\Theta}$ is the set of all parameters values that could be the population value.

- As an example, $m(\mathbf{x},\boldsymbol{\theta}) = \exp(\mathbf{x}\boldsymbol{\theta}) = \exp(\theta_1 + \theta_2 x_2 +\ldots+\theta_K x_K)$ where $\mathbf{x} = (1,x_2,\ldots,x_K)$ contains unity for convenience. The parameter space is probably $\boldsymbol{\Theta} = \mathbb{R}^K$ because it is unlikely we would restrict it ahead of time.

- We can have more of fewer parameters than covariates. For example, if

$$m(\mathbf{x},\boldsymbol{\theta}) = \exp[\mathbf{x}\boldsymbol{\beta} + \delta_1(\mathbf{x}\boldsymbol{\beta})^2 + \delta_2(\mathbf{x}\boldsymbol{\beta})^3]$$

then $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta_1, \delta_2)'$.

- If $0 \leq y \leq 1$ – sometimes called a *fractional response* – a sensible model is

$$m(\mathbf{x}, \boldsymbol{\theta}) = \frac{\exp(\mathbf{x}\boldsymbol{\theta})}{1 + \exp(\mathbf{x}\boldsymbol{\theta})} \equiv \Lambda(\mathbf{x}\boldsymbol{\theta}).$$

- For much of our development, we assume the model *correctly specified* for the conditional mean.

ASSUMPTION NLS.1: For some $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$,

$$E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o). \ \square$$

- Remember, $\boldsymbol{\theta}_o$ is just the $P \times 1$ vector of numbers we are trying to learn about. Sometimes, $\boldsymbol{\theta}_o$ is called the "true value of the parameters."
- It seems almost certain that models we use are misspecified. We will discuss that situation later.

- For some purposes, it is useful to write the equation in error form:

$$\begin{aligned} y &= m(\mathbf{x}, \boldsymbol{\theta}_o) + u \\ E(u|\mathbf{x}) &= 0, \end{aligned}$$

where the zero conditional mean holds by construction.

- Generally, other features of $D(u|\mathbf{x})$ are unrestricted. For example, if $y \geq 0$ then $u \geq -m(\mathbf{x}, \boldsymbol{\theta}_o)$. If $0 \leq y \leq 1$, then we must also have $u \leq 1 - m(\mathbf{x}, \boldsymbol{\theta}_o)$.
- Generally, we should avoid thinking of situations where u is independent of $\mathbf{x}$, and we should not even think $Var(u|\mathbf{x}) = Var(u)$.

- Assuming a correctly specified model, and the availability of a random sample, how should we estimate $\boldsymbol{\theta}_o$? It helps to know an optimization feature of a conditional mean. Generally, let $E(y|\mathbf{x}) = \mu_o(\mathbf{x})$. Assume $E(y^2) < \infty$. Then among all functions $\mu(\mathbf{x})$ with $E[\mu(\mathbf{x})^2] < \infty$,

$$E\{[y - \mu_o(\mathbf{x})]^2\} \leq E\{[y - \mu(\mathbf{x})]^2\}.$$

That is, the conditional mean is the minimum mean square predictor of y.

- Therefore, if $m(\mathbf{x}, \boldsymbol{\theta})$ is a correctly specified model of $E(y|\mathbf{x})$, then

$$E\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\} \leq E\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\}, \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

- A direct proof is constructive. Write $y = m(\mathbf{x},\boldsymbol{\theta}_o) + u$ and plug in:

$$\begin{aligned}[y - m(\mathbf{x},\boldsymbol{\theta})]^2 &= [m(\mathbf{x},\boldsymbol{\theta}_o) + u - m(\mathbf{x},\boldsymbol{\theta})]^2 \\ &= u^2 + 2[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]u + [m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2\end{aligned}$$

Then

$$\begin{aligned}E\{[y - m(\mathbf{x},\boldsymbol{\theta})]^2\} &= E(u^2) + E\{2[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]u\} \\ &\quad + E\{[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2\} \\ &= E(u^2) + E\{[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2\}\end{aligned}$$

because $E(u|\mathbf{x}) = 0$.

- Now $E(u^2)$ does not depend on $\boldsymbol{\theta}$ and $E\{[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2\}$ is smallest when $\boldsymbol{\theta} = \boldsymbol{\theta}_o$.

- So, we have shown that

$$\boldsymbol{\theta}_o = \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\{[y - m(\mathbf{x},\boldsymbol{\theta})]^2\}.$$

- In other words, $\boldsymbol{\theta}_o$ solves a population minimization problem.

- The *analogy principle* says to solve the sample analog of the population problem, which leads to

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N}[y_i - m(\mathbf{x},\boldsymbol{\theta})]^2.$$

- The M-estimation principle generalizes this reasoning. We assume that $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$ *uniquely* solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} E[q(\mathbf{w}, \boldsymbol{\theta})]$$

where $q : \mathcal{W} \times \boldsymbol{\Theta} \to \mathbb{R}$ is a real valued function of an observable vector $\mathbf{w}$ and the parameter vector $\boldsymbol{\theta}$.

- An M-estimator of $\boldsymbol{\theta}_o$ solves the sample analog,

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}).$$

- Does it seem reasonable that a solution, say $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$ is consistent for $\boldsymbol{\theta}_o$?

- By the law of large numbers, for each $\boldsymbol{\theta}$,

$$\underbrace{N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta})}_{\substack{\hat{\boldsymbol{\theta}} \text{ minimizes} \\ \text{(sample average)}}} \xrightarrow{p} \underbrace{E[q(\mathbf{w},\boldsymbol{\theta})]}_{\substack{\boldsymbol{\theta}_o \text{ minimizes} \\ \text{(population average)}}}$$

So $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$ (as $N \to \infty$, as always) seems reasonable.

- But pointwise convergence of the sample objective function is not sufficient for consistency. A sufficient condition is *uniform convergence in probability*:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta}) - E[q(\mathbf{w},\boldsymbol{\theta})] \right| \overset{p}{\rightarrow} 0$$

- Means that we can bound the distance between $N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta})$ and its expected value by something that does not depend on $\boldsymbol{\theta}$.
- In “regular” cases, the pointwise law of large numbers translates into the *uniform law of large numbers*. Sufficient is that $q(\mathbf{w},\cdot)$ is continuous on $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}$ is closed and bounded (compact), and $|q(\mathbf{w},\boldsymbol{\theta})| \leq b(\mathbf{w})$ for some function $b(\mathbf{w})$ with $E[b(\mathbf{w})] < \infty$.

- Other than "regularity conditions" – continuity of $q(\mathbf{w},\bullet)$ and finite moments – the key consistency assumption is identification. Namely, $\boldsymbol{\theta}_o$ is the *unique* solution to the population problem.

EXAMPLE: Suppose $x > 0$ is a scalar, $y \geq 0$, and $m(x,\theta_1,\theta_2,\theta_3) = \theta_1 + \theta_2 x^{\theta_3}$, where the parameter spaces is $[0,\infty) \times [0,\infty) \times \mathbb{R}$. If $\theta_{o2} = 0$, so that $E(y|x) = E(y) = \theta_{o1}$, then any $\boldsymbol{\theta}$ of the form $(\theta_{o1},0,\theta_3)'$ minimizes the mean squared error. What would NLS estimate for θ_{o3}?

- For NLS, we can write the identification as

ASSUMPTION NLS.2: $E\{[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2\} > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$. □

- Assumption NLS.2 plays the role of the rank condition. In the linear case, $m(\mathbf{x},\boldsymbol{\theta}) = \mathbf{x}\boldsymbol{\theta}$, and then

$$m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta}) = [(\boldsymbol{\theta}_o - \boldsymbol{\theta})\mathbf{x}]^2 = (\boldsymbol{\theta}_o - \boldsymbol{\theta})'\mathbf{x}'\mathbf{x}(\boldsymbol{\theta}_o - \boldsymbol{\theta})$$
$$E\{[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2\} = (\boldsymbol{\theta}_o - \boldsymbol{\theta})'E(\mathbf{x}'\mathbf{x})(\boldsymbol{\theta}_o - \boldsymbol{\theta})$$

For the last expression to be positive for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$, we need $E(\mathbf{x}'\mathbf{x})$ to have full rank K, which is exactly Assumption OLS.2.

- Theorem 12.2 contains a formal consistency result for general M-estimators. Practically important restriction is continuity of $q(\mathbf{w},\bullet)$. Can be easily relaxed to “continuity with probability one.”
- When $q(\mathbf{w},\bullet)$ is continuous on $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ is compact, there is always a solution to

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta}).$$

It need not be unique – least absolute deviations is sometimes not unique – even though the solution to the population problem is unique.

- Very useful result (Lemma 12.1): Under finite moment conditions, if $r(\mathbf{w}, \cdot)$ is continuous on $\boldsymbol{\Theta}$,

$$N^{-1} \sum_{i=1}^{N} r(\mathbf{w}_i, \boldsymbol{\theta}) \xrightarrow{p} E[r(\mathbf{w}, \boldsymbol{\theta})], \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

and $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_o$, then

$$N^{-1} \sum_{i=1}^{N} r(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) \xrightarrow{p} E[r(\mathbf{w}, \boldsymbol{\theta}_o)]$$

- This result is useful for estimating average partial effects (which we will cover later) as well as asymptotic variance matrices.

3. ASYMPTOTIC DISTRIBUTION

- In the previous section, we showed how consistency of M-estimators can be established without having closed form solutions. Now we turn to the question of approximating the sampling distribution of $\hat{\boldsymbol{\theta}}$.
- We now add some smoothness assumptions. In particular, assume

$$q(\mathbf{w},\cdot) \text{ is twice continuously differentiable on } int(\boldsymbol{\Theta}).$$

- Further, assume $\boldsymbol{\theta}_o$ is in the interior of the parameter space:

$$\boldsymbol{\theta}_o \in int(\boldsymbol{\Theta}).$$

- Finite moment conditions are used, too.

- The gradient of $q(\mathbf{w}, \boldsymbol{\theta})$, defined on $int(\boldsymbol{\Theta})$, is the $1 \times P$ row vector

$$\nabla_{\boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}) = \left(\begin{array}{cccc} \frac{\partial q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_P} \end{array} \right).$$

The *score* is the transpose of the gradient:

$$\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta})'.$$

- Now, because $\boldsymbol{\theta}_o$ is in the interior of $\boldsymbol{\Theta}$ and $\hat{\boldsymbol{\theta}} \overset{p}{\rightarrow} \boldsymbol{\theta}_o$, we know $\hat{\boldsymbol{\theta}} \in int(\boldsymbol{\Theta})$ *with probability approaching one*. We will ignore the qualifier here.

- Because $\hat{\boldsymbol{\theta}}$ minimizes the sample objective function and is an interior solution, $\hat{\boldsymbol{\theta}}$ solves

$$\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0},$$

a set of P equations in P unknowns. (Many algorithms to actually find $\hat{\boldsymbol{\theta}}$ are based on this first order condition.) Because $q(\mathbf{w}, \bullet)$ is twice continuously differentiable, each $s_m(\mathbf{w}, \bullet)$, $m = 1, \ldots, P$, is continuously differentiable.

- By the mean value theorem (for each element of the score),

$$\sum_{i=1}^{N} s_m(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} s_m(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} s_m(\mathbf{w}_i, \ddot{\boldsymbol{\theta}}_m) \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$$

where $\ddot{\boldsymbol{\theta}}_m$ is on the line segment between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$ for $m = 1, \ldots, P$. Therefore, $\ddot{\boldsymbol{\theta}}_m \overset{p}{\rightarrow} \boldsymbol{\theta}_o$. (In effect, $\ddot{\boldsymbol{\theta}}_m$ is “trapped” beween $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$.)

- Stack all P elements to get

$$\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o),$$

where $\ddot{\mathbf{H}}_i$ is the $P \times P$ Hessian of $q(\mathbf{w}, \boldsymbol{\theta})$ – also the Jacobian of $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$ – but with rows evaluated at generally different mean values.

- We will need the Hessian evaluated at a generic parameter vector:

$$\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) = \frac{\partial^2 q(\mathbf{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

- Back to the score representation. Because $\hat{\boldsymbol{\theta}}$ solves the FOC,

$$\mathbf{0} = \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$$

so

$$\mathbf{0} = N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right) \sqrt{N} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o).$$

- Because each $\ddot{\boldsymbol{\theta}}_m \overset{p}{\to} \boldsymbol{\theta}_o$, $N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \overset{p}{\to} E[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)] \equiv \mathbf{A}(\boldsymbol{\theta}_o) \equiv \mathbf{A}_o$ by Lemma 12.1

- An assumption related to identification is that

$$\mathbf{A}_o \text{ is positive definite}$$

- Can show that $N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i$ is nonsingular w.p.a.1. because it is getting “close” to $\mathbf{A}_o$, which is nonsingular.

- It follows that, w.p.a.1.,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right)^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) \right]$$

- Further, very generally the score has zero mean when evaluated at $\boldsymbol{\theta}_o$:

$$E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}.$$

(Important: $E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})] \neq \mathbf{0}$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$.)

- Can show $E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}$ if the derivative and expected value can be interchanged. By FOC in the population,

$$\nabla_{\boldsymbol{\theta}} E[q(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0},$$

and so $E[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}$ if $\nabla_{\boldsymbol{\theta}}$ can pass through the expected value. (This is shown generally in some analysis books.)

- Similar argument can be used to show $\mathbf{A}_o$ is positive definite. Why? Because $\boldsymbol{\theta}_o$ uniquely solves the population minimization problem,

$$\nabla^2_{\boldsymbol{\theta}} E[q(\mathbf{w}, \boldsymbol{\theta}_o)]$$

is positive definite. Now interchange the partial derivatives with respect to $\boldsymbol{\theta}$ and the expected value and we get $\mathbf{A}_o$.

- In other cases, such as NLS, can show directly the score has zero expected value at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$.
- Why is $E[\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)] = \mathbf{0}$ important? Because then, by the central limit theorem,

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o) \overset{d}{\to} Normal(\mathbf{0},\mathbf{B}_o)$$

$$\mathbf{B}_o = Var[\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o)] = E[\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o)'].$$

- In particular, $N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o) = O_p(1)$.

- Now,

$$
\begin{aligned}
\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) &= \mathbf{A}_o^{-1}\left[-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) \right] \\
&\quad + \left[\left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right)^{-1} - \mathbf{A}_o^{-1} \right] \left[-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) \right] \\
&= \mathbf{A}_o^{-1}\left[-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) \right] + o_p(1) \cdot O_p(1) \\
&= \mathbf{A}_o^{-1}\left[-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) \right] + o_p(1).
\end{aligned}
$$

- If we define $\mathbf{r}_i(\boldsymbol{\theta}_o) = -\mathbf{A}_o^{-1}\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)$, then we can write

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = N^{-1/2}\sum_{i=1}^{N}\mathbf{r}_i(\boldsymbol{\theta}_o) + o_p(1),$$

which is called the *influence function representation.*

- Notice that $E[\mathbf{r}_i(\boldsymbol{\theta}_o)] = \mathbf{0}$ and

$Var[\mathbf{r}_i(\boldsymbol{\theta}_o)] = \mathbf{A}_o^{-1}Var[\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)]\mathbf{A}_o^{-1} = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$.

- By the asymptotic equivalence lemma,

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \overset{d}{\to} Normal(\mathbf{0}, \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}).$$

- Generally, the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ depends on the expected value of the Hessian and the variance of the score (both evaluated at $\boldsymbol{\theta}_o$).

- The expression for $Avar[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]$ is of the “sandwich” form (although in some cases it simplifies).
- We write

$$Avar(\hat{\boldsymbol{\theta}}) = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}/N,$$

so that $\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}/N$ is intended to approximate the actual sampling variation in $\hat{\boldsymbol{\theta}}$ for a give sample size, N.

- Note that, as with simple estimators, such as sample averages, $Avar(\hat{\boldsymbol{\theta}})$ is of order $1/N$.

4. ESTIMATING THE ASYMPTOTIC VARIANCE

- Technically, we must talk about consistent estimation of $Avar[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]$, as this is the quantity that does not depend on N. So we must consistently estimate $\mathbf{A}_o$ and $\mathbf{B}_o$.
- There are sometimes several different ways to estimate $\mathbf{A}_o$. An estimator that is always available is simply

$$N^{-1}\sum_{i=1}^{N}\mathbf{H}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = N^{-1}\sum_{i=1}^{N}\mathbf{H}_i(\hat{\boldsymbol{\theta}}),$$

the average of the Hessians evaluated at the estimates.

- When $\mathbf{w}_i$ partitions as $(\mathbf{x}_i, \mathbf{y}_i)$, and we are correctly modeling a feature of $D(\mathbf{y}_i|\mathbf{x}_i)$, we can often find

$$\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = E[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)|\mathbf{x}_i].$$

By iterated expectations, $\mathbf{A}_o = E[\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o)]$. So a second consistent estimator of $\mathbf{A}_o$ is sometimes available:

$$N^{-1}\sum_{i=1}^{N}\mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) = N^{-1}\sum_{i=1}^{N}\hat{\mathbf{A}}_i.$$

- This is the estimator based on the “expected Hessian,” although emphasizing the conditioning on $\mathbf{x}$ is more precise.

- It is rarely possible to find the unconditional expected value of $\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)$ when there are conditioning variables because we are not usually modeling $D(\mathbf{x}_i)$.
- A natural consistent estimator of $\mathbf{B}_o = E[\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)']$ is

$$\begin{aligned}\hat{\mathbf{B}} &= N^{-1}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})\mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})' = N^{-1}\sum_{i=1}^{N}\mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})' \\ &= N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\end{aligned}$$

- Called the *outer product of the score*.

- Therefore,

$$\begin{aligned}\widehat{Avar}(\hat{\boldsymbol{\theta}}) &= N^{-1}\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\right)\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\\ &= \left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\right)\left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\end{aligned}$$

- As with all other procedures, the divions by N disappear in $\widehat{Avar}(\hat{\boldsymbol{\theta}})$.

- If we can compute $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = E[\mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}_o)|\mathbf{x}_i]$ then we can use

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{N} \hat{\mathbf{A}}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i'\right) \left(\sum_{i=1}^{N} \hat{\mathbf{A}}_i\right)^{-1}$$

- When the inverses exist, both estimators are always at least positive semidefinite, and usually positive definite unless the underlying model is poorly specified.

Nonlinear Least Squares

- Write $q(\mathbf{w},\boldsymbol{\theta}) = [y - m(\mathbf{x},\boldsymbol{\theta})]^2/2$. Then

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} q(\mathbf{w},\boldsymbol{\theta}) &= -\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta})[y - m(\mathbf{x},\boldsymbol{\theta})] \\ \mathbf{s}(\mathbf{w},\boldsymbol{\theta}) &= -\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta})'[y - m(\mathbf{x},\boldsymbol{\theta})]\end{aligned}$$

assuming $m(\mathbf{x},\cdot)$ is twice continuously differentiable on $int(\boldsymbol{\Theta})$.

- The NLS estimator satisfies $\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i,\hat{\boldsymbol{\theta}})'[y_i - m(\mathbf{x}_i,\hat{\boldsymbol{\theta}})] = \mathbf{0}$. In the linear case, $m(\mathbf{x}_i,\hat{\boldsymbol{\theta}}) = \mathbf{x}_i\hat{\boldsymbol{\theta}}$ and $\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i,\hat{\boldsymbol{\theta}}) = \mathbf{x}_i$, and we obtain the FOC for the OLS estimator.

- We can directly show the conditional mean of the score is zero at $\boldsymbol{\theta}_o$:

$$\begin{aligned} E[\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)|\mathbf{x}] &= -\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)'[E(y|\mathbf{x}) - m(\mathbf{x},\boldsymbol{\theta}_o)] \\ &= -\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)' \cdot 0 = \mathbf{0}. \end{aligned}$$

(In fact, can write $\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o) = -\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)'u$ where $u \equiv y - m(\mathbf{x},\boldsymbol{\theta}_o)$.)

- By iterated expectations, of course, $E[\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)] = \mathbf{0}$.

- The Hessian is

$$\mathbf{H}(\mathbf{w},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}(\mathbf{w},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta})'\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}^2 m(\mathbf{x},\boldsymbol{\theta})u(\boldsymbol{\theta})$$

where $u(\boldsymbol{\theta}) \equiv y - m(\mathbf{x},\boldsymbol{\theta})$. Note that $\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta})'\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}}^2 m(\mathbf{x},\boldsymbol{\theta})$ are $P \times P$.

- We can use $\mathbf{H}(\mathbf{w},\boldsymbol{\theta})$ to estimate the asymptotic variance, or use $E(u|\mathbf{x}) = 0$:

$$\begin{aligned}\mathbf{A}(\mathbf{x},\boldsymbol{\theta}_o) &= E[\mathbf{H}(\mathbf{w},\boldsymbol{\theta}_o)|\mathbf{x}] = \nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)'\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o) - \nabla^2_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)E[u(\boldsymbol{\theta}_o)|\mathbf{x}] \\ &= \nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)'\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o).\end{aligned}$$

- A typical estimator of $\mathbf{A}_o$ for NLS is

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i,\hat{\boldsymbol{\theta}})'\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i,\hat{\boldsymbol{\theta}}) = N^{-1}\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\hat{m}_i'\nabla_{\boldsymbol{\theta}}\hat{m}_i,$$

but it does assume correct specification of the conditional mean. (In linear case, reduces to the usual estimator no matter what.)

- For $\mathbf{B}_o$, $\hat{\mathbf{s}}_i = -\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})'[y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})] \equiv -\nabla_{\boldsymbol{\theta}} \hat{m}_i' \hat{u}_i$ where $\hat{u}_i \equiv y_i - m(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ are the NLS residuals.

$$\hat{\mathbf{B}} = N^{-1} \sum_{i=1}^{N} \hat{\mathbf{s}}_i' \hat{\mathbf{s}}_i = N^{-1} \sum_{i=1}^{N} \hat{u}_i^2 \nabla_{\boldsymbol{\theta}} \hat{m}_i' \nabla_{\boldsymbol{\theta}} \hat{m}_i$$

- Combining gives the Huber-White heteroskedasticity-robust variance matrix estimator:

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \hat{m}_i' \nabla_{\boldsymbol{\theta}} \hat{m}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{u}_i^2 \nabla_{\boldsymbol{\theta}} \hat{m}_i' \nabla_{\boldsymbol{\theta}} \hat{m}_i \right) \left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \hat{m}_i' \nabla_{\boldsymbol{\theta}} \hat{m}_i \right)^{-1}$$

(under correct specification of the conditional mean).

- Suppose we add a homoskedasticity assumption:

ASSUMPTION NLS.3: $Var(y|\mathbf{x}) = \sigma_o^2$. □

- With Assumption NLS.3, the expression simplifies, just as with OLS:

$$\begin{aligned} E[\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)'|\mathbf{x}] &= E[u^2\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta}_o)'\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta}_o)|\mathbf{x}] \\ &= E(u^2|\mathbf{x})\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta}_o)'\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta}_o) \\ &= \sigma_o^2\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta}_o)'\nabla_{\boldsymbol{\theta}}m(\mathbf{x},\boldsymbol{\theta}_o) \end{aligned}$$

because $E(u^2|\mathbf{x}) = Var(u|\mathbf{x}) = Var(y|\mathbf{x})$ when $E(u|\mathbf{x}) = 0$.

- Aside: NLS.3 does *not* say that $Var(y|\mathbf{x}) = Var(y)$.

- So

$$\begin{aligned}\mathbf{B}_o &= E[\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)'] = \sigma_o^2 E[\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)'\nabla_{\boldsymbol{\theta}} m(\mathbf{x},\boldsymbol{\theta}_o)] \\ &= \sigma_o^2 \mathbf{A}_o.\end{aligned}$$

- A consistent estimator of σ_o^2 (with a degree-of-freedom adjustment) is

$$\hat{\sigma}^2 = (N-P)^{-1}\sum_{i=1}^{N}\hat{u}_i^2$$

and then

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\hat{m}_i'\nabla_{\boldsymbol{\theta}}\hat{m}_i\right)^{-1}.$$

5. LARGE-SAMPLE INFERENCE

- The asymptotic standard errors are obtained as the square roots of the diagonal elements.

$$\frac{(\hat{\theta}_j - \theta_{oj})}{se(\hat{\theta}_j)} \overset{d}{\to} Normal(0,1).$$

- Therefore, to test $H_0 : \theta_{oj} = a_j$, use

$$t(\hat{\theta}_j, a_j) = \frac{(\hat{\theta}_j - a_j)}{se(\hat{\theta}_j)}$$

as approximately $Normal(0,1)$. Obtain approximate confidence intervals, too: $\hat{\theta}_j \pm 1.96 \cdot se(\hat{\theta}_j)$ for a large-sample 95% CI.

- We can test multiple, nonlinear restrictions. Let H_0 be stated as

$$H_0 : \mathbf{c}(\boldsymbol{\theta}_o) = \mathbf{0}$$

where $\mathbf{c} : \boldsymbol{\Theta} \to \mathbb{R}^Q$, so there are Q restrictions. Assume $\mathbf{c}(\cdot)$ is continuously differentiable on $int(\boldsymbol{\Theta})$ and that $\boldsymbol{\theta}_o \in int(\boldsymbol{\Theta})$, as before. Let $\mathbf{C}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{c}(\boldsymbol{\theta})$ be the $Q \times P$ Jacobian, and define $\mathbf{C}_o \equiv \mathbf{C}(\boldsymbol{\theta}_o)$.

- By the mean value theorem argument,

$$\begin{aligned}\sqrt{N}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{c}(\boldsymbol{\theta}_o)] &= \nabla_{\boldsymbol{\theta}}\mathbf{c}(\boldsymbol{\theta}_o)\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1) \\ &= \mathbf{C}_o\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) + o_p(1).\end{aligned}$$

- So, under H_0,

$$\sqrt{N}\,\mathbf{c}(\hat{\boldsymbol{\theta}}) \overset{d}{\rightarrow} Normal(\mathbf{0}, \mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\mathbf{C}_o')$$

$$\sqrt{N}\,\mathbf{c}(\hat{\boldsymbol{\theta}})'[\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\mathbf{C}_o']^{-1}\sqrt{N}\,\mathbf{c}(\hat{\boldsymbol{\theta}}) \overset{d}{\rightarrow} \chi_Q^2$$

- The *Wald statistic* is

$$\begin{aligned} W &= N\mathbf{c}(\hat{\boldsymbol{\theta}})'(\hat{\mathbf{C}}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{\mathbf{-1}}\hat{\mathbf{C}}')^{-1}\mathbf{c}(\hat{\boldsymbol{\theta}}) \\ &= \mathbf{c}(\hat{\boldsymbol{\theta}})'\{\widehat{Avar}[\mathbf{c}(\hat{\boldsymbol{\theta}})]\}^{-1}\mathbf{c}(\hat{\boldsymbol{\theta}}) \end{aligned}$$

where $\hat{\mathbf{C}} = \mathbf{C}(\hat{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}}\mathbf{c}(\hat{\boldsymbol{\theta}})$.

- Under H_0,

$$W \overset{d}{\to} \chi^2_Q.$$

- The Wald statistic is convenient when the unrestricted model is easy to estimate. It is almost always available, and can be made robust by using the sandwich form of the asymptotic variance estimate.
- Typically, the Wald statistic is the default when reported by econometrics packages.

- The *score statistic* or *Lagrange multiplier statistic* is based only on the restricted estimate. Let $\tilde{\boldsymbol{\theta}}$ be the estimator that solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta})$$

$$\text{subject to } \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$$

- The score principle is to insert the restricted estimates into the unrestricted score, and then seeing "how far" the result is from zero.

- Based on a mean value expansion, can show under H_0 that

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\tilde{\boldsymbol{\theta}}) = N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o) + \mathbf{A}_o\sqrt{N}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_o) + o_p(1).$$

- By another mean value expansion,

$$\sqrt{N}\,\mathbf{c}(\tilde{\boldsymbol{\theta}}) = \sqrt{N}\,\mathbf{c}(\boldsymbol{\theta}_o) + \mathbf{C}_o\sqrt{N}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_o) + o_p(1).$$

But $\mathbf{c}(\boldsymbol{\theta}_o) = \mathbf{0}$ under H_0 and $\mathbf{c}(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$ because $\tilde{\boldsymbol{\theta}}$ is the restricted estimator. So $\mathbf{C}_o\sqrt{N}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_o) = o_p(1)$.

- Theforefore, under H_0,

$$\mathbf{C}_o\mathbf{A}_o^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\tilde{\boldsymbol{\theta}})\right) = \mathbf{C}_o\mathbf{A}_o^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o)\right) + o_p(1)$$

so

$$\mathbf{C}_o\mathbf{A}_o^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\tilde{\boldsymbol{\theta}})\right) \overset{d}{\to} Normal(\mathbf{0},\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}\mathbf{C}_o')$$

- The LM statistic is a quadratic form in $N^{-1/2}\sum_{i=1}^{N}\tilde{\mathbf{s}}_i$:

$$LM = \left(\sum_{i=1}^{N}\tilde{\mathbf{s}}_i\right)'\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{C}}'(\tilde{\mathbf{C}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{C}}')^{-1}\tilde{\mathbf{C}}'\tilde{\mathbf{A}}^{-1}\left(\sum_{i=1}^{N}\tilde{\mathbf{s}}_i\right)/N \overset{d}{\rightarrow} \chi^2_Q.$$

- All quantities are evaluated at $\tilde{\boldsymbol{\theta}}$. For example,

$$\tilde{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\tilde{\mathbf{s}}_i\tilde{\mathbf{s}}_i'$$

- Depending on the choice of $\tilde{\mathbf{A}}$, it may not be positive definite (particularly the Hessian form). But the LM statistic above is always nonnegative.
- This is a fully robust LM statistic in that it only assumes $\boldsymbol{\theta}_o$ minimizes $E[q(\mathbf{w}_i, \boldsymbol{\theta}_o)]$ subject to $\mathbf{c}(\boldsymbol{\theta}_o) = \mathbf{0}$ (and, the regularity conditions of $\boldsymbol{\theta}_o$ being in the interior under H_0, differentiability, and moment conditions).

- If the *generalized information matrix equality*, that is, for some $\sigma_o^2 > 0$,

$$\mathbf{B}_o = \sigma_o^2 \mathbf{A}_o$$

then the LM statistic simplifies to

$$LM = \left(\sum_{i=1}^{N} \tilde{\mathbf{s}}_i\right)' \tilde{\mathbf{M}}^{-1} \left(\sum_{i=1}^{N} \tilde{\mathbf{s}}_i\right) / \tilde{\sigma}^2,$$

where $\tilde{\sigma}^2$ is a consistent estimator of σ_o^2.

- The matrix $\tilde{\mathbf{M}}$ can be one of the matrices

$$\sum_{i=1}^{N} \tilde{\mathbf{H}}_i, \ \sum_{i=1}^{N} \tilde{\mathbf{A}}_i, \ \sum_{i=1}^{N} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i'.$$

- These lead to the "Hessian," "expected Hessian," and "outer product" of the LM or score statistics. This is a nonrobust statistic because it uses the GIME.

- When $\sigma_o^2 = 1$ (as in correctly specified maximum likelihood, as we will see), the outer product statistic is $N - SSR_0 = NR_0^2$ from the regression (without a constant),

$$1 \text{ on } \tilde{\mathbf{s}}_i', \; i = 1, \ldots, N.$$

(R_o^2 is the "uncentered" R-squared, that is, the dependent variable is not centered about its sample mean.)

- The nonrobust form of the statistic need not be positive if $\sum_{i=1}^{N} \tilde{\mathbf{H}}_i$ is used because the Hessian at the restricted estimates need not be positive definite. Usually the estimated expected Hessian is positive definite (more later), and the outer product form is always nonnegative.
- The outer product form, while computationally simple, has been shown sometimes to have serious size distortions even in pretty large samples.
- In some leading cases, the expected Hessian depends only on first derivatives and is always nonnegative.

EXAMPLE: Nonlinear Least Squares. Suppose that $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\delta}')'$ where $\boldsymbol{\delta}$ is $Q \times 1$, and the mean is correctly specified. State $H_0 : \boldsymbol{\delta}_o = \bar{\boldsymbol{\delta}}$ for a specified set of values $\bar{\boldsymbol{\delta}}$. Let $\tilde{\boldsymbol{\beta}}$ be the NLS estimator subject to $\boldsymbol{\delta} = \bar{\boldsymbol{\delta}}$. Then $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \bar{\boldsymbol{\delta}}')'$ and the $1 \times P$ gradient of the mean function is

$$\nabla_{\boldsymbol{\theta}} m(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}) = [\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \tilde{\boldsymbol{\theta}}), \nabla_{\boldsymbol{\delta}} m(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})].$$

● If we make NLS.3 (homoskedasticty) under the null, the expected Hessian form of the statistic is NR_u^2 from the auxiliary regression

$$\tilde{u}_i \text{ on } \nabla_{\boldsymbol{\beta}} \tilde{m}_i, \nabla_{\boldsymbol{\delta}} \tilde{m}_i,\ i = 1, \ldots, N$$

where $\tilde{u}_i = y_i - m(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})$ are the restricted NLS residuals and R_u^2 is the uncentered R-squared.

- Notice that even though $\sum_i^N \nabla_{\boldsymbol{\beta}} \tilde{m}_i' \tilde{u}_i = \mathbf{0}$ by the FOC for the constrained NLS estimator, $\nabla_{\boldsymbol{\beta}} \tilde{m}_i$ needs to be included because it is generally correlated with $\nabla_{\boldsymbol{\delta}} \tilde{m}_i$.
- A regression form of the robust test: (1) Regress $\nabla_{\boldsymbol{\delta}} \tilde{m}_i$ on $\nabla_{\boldsymbol{\beta}} \tilde{m}_i$ and obtain the $1 \times Q$ residuals, $\tilde{\mathbf{r}}_i$. (2) Use the usual heteroskedastic-robust Wald statistic of joint significance in the regression $\tilde{u}_i$ on $\tilde{\mathbf{r}}_i$, $i = 1,\ldots,N$. (Or, use $N - SSR_0 = NR_0^2$ from the regression 1 on $\tilde{u}_i \tilde{\mathbf{r}}_i$, $i = 1,\ldots,N$.)

- Under the GIME, a statistic based on the change in the criterion function is available. This requires estimation of both the restricted and unrestricted models but no matrix algebra for computation.
- We know that

$$\sum_{i=1}^{N} q(\mathbf{w}_i, \tilde{\boldsymbol{\theta}}) \geq \sum_{i=1}^{N} q(\mathbf{w}_i, \hat{\boldsymbol{\theta}})$$

because $\tilde{\boldsymbol{\theta}}$ is the restricted estimator.

- Define the *quasi-likelihood ratio (QLR) statistic* as

$$QLR = 2\left[\sum_{i=1}^{N} q(\mathbf{w}_i, \tilde{\boldsymbol{\theta}}) - \sum_{i=1}^{N} q(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) \right] / \hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is a consistent estimator of σ_o^2 typically obtained without the restrictions imposed.

- Under $H_0 : \mathbf{c}(\boldsymbol{\theta}_o) = \mathbf{0}$ and $\mathbf{B}_o = \sigma_o^2 \mathbf{A}_o$,

$$QLR \overset{d}{\rightarrow} \chi_Q^2$$

- The QLR name comes from its application to maximum likelihood estimation. But it is also related to the usual F statistic from linear regression. In fact, if $q(\mathbf{w}_i, \boldsymbol{\theta}) = (y_i - \mathbf{x}_i\boldsymbol{\theta})^2/2$ and $\hat{\sigma}^2 = (N - P)^{-1} \sum_{i=1}^{N} \hat{u}_i^2$, then $QLR = Q \cdot F$, where F is the usual F statistic for testing the Q restrictions.

- Generally, for NLS under NLS.3, we can use

$$F = \frac{(SSR_r - SSR_{ur})}{SSR_{ur}} \cdot \frac{(N - P)}{Q}$$

as approximately $\mathcal{F}_{Q,N-P}$ under H_0. No theoretical reason not to use QLR as approximate χ^2_Q, but the "F" statistic has been shown to sometimes have better size in not-so-large samples.

Choosing Among the Statistics

- Under the null, the robust versions of the Wald and LM statistics have the same limiting chi-square distribution. The QLR statistic does under the GIME, and, of course, there are nonrobust versions of the Wald and LM statistics.
- Can we use power considerations to choose among the statistics? Against fixed alternatives, all three statistics reject with probability approaching one. That is, they are *consistent tests.*

- Instead, use a *local alternatives* approach. Suppose the sequence of "true" parameters is $\{\boldsymbol{\theta}_{o,N} : N = 1, 2, \ldots\}$ and these satisfy

$$\mathbf{c}(\boldsymbol{\theta}_{o,N}) = \boldsymbol{\delta}_o/\sqrt{N}$$

for some $Q \times 1$ vector $\boldsymbol{\delta}_o$. So, the null is violated for each N, but it is closer to being true as the sample size grows. We can derive the *asymptotic local power* of each of the statistics. The Wald statistic is easiest to study. Let $\boldsymbol{\theta}_o$ denote the limit of $\boldsymbol{\theta}_{o,N}$. Then a mean value expansion gives

$$\sqrt{N}\,\mathbf{c}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\delta}_o + \mathbf{C}_o\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{o,N}) + o_p(1).$$

- Assume the GIME for simplicity. Then

$$\sqrt{N}\,\mathbf{c}(\hat{\boldsymbol{\theta}}) \overset{d}{\rightarrow} Normal(\boldsymbol{\delta}_o, \mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{C}_o')$$

and so the Wald statistic has a limting chi-square distribution with *noncentrality parameter*

$$\boldsymbol{\delta}_o'(\mathbf{C}_o\mathbf{A}_o^{-1}\mathbf{C}_o')^{-1}\boldsymbol{\delta}_o.$$

- Turns out to be the same for the LM and QLR statistics, so we cannot choose among the tests based on local power analysis.

- Given $\boldsymbol{\delta}_o$, we can estimate the noncentrality parameter as $\boldsymbol{\delta}_o'(\hat{\mathbf{C}}\hat{\mathbf{A}}^{-1}\hat{\mathbf{C}}')^{-1}\boldsymbol{\delta}_o$ and then do (local) power analysis.
- Same conclusions for Wald and LM when we look at the robust versions of the statistics.
- Typically, the choice is based on computational simplicity and evidence of finite-sample peformance.
- Local power analysis is useful when comparing different estimation methods. A more (asymptotically) efficient estimator leads to a larger noncentrality parameter and higher local power.

6. TWO-STEP ESTIMATION

- Let $\hat{\boldsymbol{\gamma}}$ be an estimator of a set of parameters ($J \times 1$) from a preliminary estimation problem. A *two-step M-estimator* solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}; \hat{\boldsymbol{\gamma}}).$$

- Assume that $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$, where $\boldsymbol{\gamma}^*$ is a fixed set of values. We use this notation to emphasize the possibility that $\hat{\boldsymbol{\gamma}}$ comes from a "misspecified" estimation problem.

- Under regularity conditions (continuity of q in $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, finite moments) the key condition for consistency is that $\boldsymbol{\theta}_o$ uniquely solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[q(\mathbf{w}_i,\boldsymbol{\theta};\boldsymbol{\gamma}^*)].$$

- Sometimes, $\boldsymbol{\theta}_o$ solves the population problem for *any* $\boldsymbol{\gamma}$. That is, for all $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$,

$$\boldsymbol{\theta}_o = \text{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E[q(\mathbf{w}_i,\boldsymbol{\theta};\boldsymbol{\gamma})].$$

In effect, we can use any first-step estimator as long as it converges to something.

Weighted Nonlinear Least Squares

- Let $h(\mathbf{x},\boldsymbol{\gamma}) > 0$ be a model of the variance function $Var(y|\mathbf{x})$. For now, do not assume it is correctly specified. But, generally, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$ if we use standard estimation approaches, such as two-step M-estimation. If $\check{u}_i = y_i - m(\mathbf{x}_i,\check{\boldsymbol{\theta}})$,where $\check{\boldsymbol{\theta}}$ is the NLS estimate, then $\hat{\boldsymbol{\gamma}}$ might solve

$$\min_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \sum_{i=1}^{N}[\check{u}_i^2 - h(\mathbf{x}_i,\boldsymbol{\gamma})]^2.$$

There are other possibilities, too. (The normal quasi-MLE, for example.)

- The *weighted nonlinear least squares* (WNLS) estimator solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}[y_i - m(\mathbf{x}_i,\boldsymbol{\theta})]^2/h(\mathbf{x}_i,\hat{\boldsymbol{\gamma}})$$

- Generally, without more assumptions, we cannot conclude that this is “better” than usual NLS, although it can be if $h(\mathbf{x},\boldsymbol{\gamma})$ is a “good” approximation to $Var(y|\mathbf{x})$.

- Under weak regularity conditions,

$$N^{-1}\sum_{i=1}^{N}[y_i - m(\mathbf{x}_i,\boldsymbol{\theta})]^2/h(\mathbf{x}_i,\hat{\boldsymbol{\gamma}}) \overset{p}{\to} E\{[y_i - m(\mathbf{x}_i,\boldsymbol{\theta})]^2/h(\mathbf{x}_i,\boldsymbol{\gamma}^*)\}$$

- If the conditional mean is correctly specified, $\boldsymbol{\theta}_o$ solves the population problem for *any* $\boldsymbol{\gamma}$. To show this, use a stronger property of the conditional mean: $\boldsymbol{\theta}_o$ solves, for any $\mathbf{x}_i$,

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} E\{[y_i - m(\mathbf{x}_i,\boldsymbol{\theta})]^2|\mathbf{x}_i\}.$$

- Because $h(\mathbf{x}_i, \boldsymbol{\gamma}) > 0$ and is a function of $\mathbf{x}_i$,

$$\begin{aligned} E\{[y_i - m(\mathbf{x}_i, \boldsymbol{\theta}_o)]^2 / h(\mathbf{x}_i, \boldsymbol{\gamma}) | \mathbf{x}_i\} &= E\{[y_i - m(\mathbf{x}_i, \boldsymbol{\theta}_o)]^2 | \mathbf{x}_i\} / h(\mathbf{x}_i, \boldsymbol{\gamma}) \\ &\leq E\{[y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 | \mathbf{x}_i\} / h(\mathbf{x}_i, \boldsymbol{\gamma}). \end{aligned}$$

- By iterated expectations, for any $\boldsymbol{\theta}$ and any $\boldsymbol{\gamma}$,

$$E\{[y_i - m(\mathbf{x}_i, \boldsymbol{\theta}_o)]^2 / h(\mathbf{x}_i, \boldsymbol{\gamma})\} \leq E\{[y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 / h(\mathbf{x}_i, \boldsymbol{\gamma})\}.$$

- We just have to assume (or show) uniqueness of $\boldsymbol{\theta}_o$.

- It follows that WNLS identifies the conditional mean parameters for essentially *any* positive weighting function that is a function of $\mathbf{x}_i$, and possibly parameters estimated in the first stage. That weighting function can be arbitrarily misspecified for the conditional variance provided it satisfies standard regularity conditions.

- More generally, we can count on consistency of a two-step M-estimator under weak conditions, the most important being identification and continuity of the objective function over $(\boldsymbol{\theta}, \boldsymbol{\gamma})$.
- In some cases, the objective function only identifies $\boldsymbol{\theta}_o$ when the first-stage estimation problem is “correctly specified,” in which case we would write $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_o$.
- Inference is more interesting. In general, we should expect to have to adjust the asymptotic variance of $\hat{\boldsymbol{\theta}}$ to account for preliminary estimation of $\hat{\boldsymbol{\gamma}}$ (because they are obtained from the same set of data).

- Assume $\hat{\boldsymbol{\gamma}}$ has first-order representation

$$\sqrt{N}\,(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) = N^{-1/2} \sum_{i=1}^{N} \mathbf{r}_i(\boldsymbol{\gamma}^*) + o_p(1),$$

where $E[\mathbf{r}_i(\boldsymbol{\gamma}^*)] = \mathbf{0}$ and $\mathbf{r}_i(\boldsymbol{\gamma}^*)$ often takes the form $\mathbf{M}^*\mathbf{e}_i(\boldsymbol{\gamma}^*)$ for a constant matrix $\mathbf{M}^*$, and we often have to estimate $\mathbf{M}^*$.

- To obtain the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$, use mean value expansion:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{A}_o^{-1}\left(-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}_i(\boldsymbol{\theta}_o;\hat{\boldsymbol{\gamma}}) \right) + o_p(1)$$

where $\mathbf{s}_i(\boldsymbol{\theta},\boldsymbol{\gamma})$ is the $P \times 1$ is the score with respect to $\boldsymbol{\theta}$:

$$\mathbf{s}_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \nabla_{\boldsymbol{\theta}} q(\mathbf{w}_i,\boldsymbol{\theta};\boldsymbol{\gamma})'.$$

- Now use a second MV expansion:

$$
\begin{aligned}
N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\boldsymbol{\theta}_o,\hat{\boldsymbol{\gamma}}) &= N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*) \\
&\quad + E[\nabla_{\boldsymbol{\gamma}}\mathbf{s}_i(\boldsymbol{\theta}_o,\boldsymbol{\gamma}^*)]\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*) + o_p(1) \\
&= N^{-1/2}\sum_{i=1}^{N}[\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*) + \mathbf{F}_o\mathbf{r}_i] + o_p(1)
\end{aligned}
$$

where $\mathbf{F}_o = E[\nabla_{\boldsymbol{\gamma}}\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*)]$ is the $P \times J$ expected Jacobian of $\mathbf{s}_i(\boldsymbol{\theta},\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$, evaluated at $(\boldsymbol{\theta}_o,\boldsymbol{\gamma}^*)$.

- Collecting terms,

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) &= \mathbf{A}_o^{-1}\left(-N^{-1/2} \sum_{i=1}^{N} [\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*) + \mathbf{F}_o \mathbf{r}_i(\boldsymbol{\gamma}^*)] \right) + o_p(1) \\ &\equiv \mathbf{A}_o^{-1}\left(-N^{-1/2} \sum_{i=1}^{N} \mathbf{g}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*) \right) + o_p(1)\end{aligned}$$

- Let $\mathbf{D}_o = Var[\mathbf{g}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*)]$. Then

$$Avar[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] = \mathbf{A}_o^{-1}\mathbf{D}_o\mathbf{A}_o^{-1}$$

- If we ignore the estimation error in $\hat{\boldsymbol{\gamma}}$, we would use $\mathbf{B}_o = Var[\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*)]$ in place of $\mathbf{D}_o$and ignore $\mathbf{F}_o\mathbf{r}_i$
- In some cases, $\mathbf{F}_o$ is equal to zero, and so it is legitimate to ignore estimation of $\boldsymbol{\gamma}^*$.
- In other cases, $\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*)$ and $\mathbf{r}_i^*$ are uncorrelated, in which case the variance that ignores $\hat{\boldsymbol{\gamma}}$ is too small:

$$\begin{aligned}\mathbf{D}_o &= \mathbf{B}_o + \mathbf{F}_o\mathbf{L}^*\mathbf{F}_o' \\ \mathbf{L}^* &\equiv E[\mathbf{r}_i(\boldsymbol{\gamma}^*)\mathbf{r}_i(\boldsymbol{\gamma}^*)']\end{aligned}$$

- There are even cases where $\mathbf{B}_o - \mathbf{D}_o$ is actually positive semidefinite, in which case the correct formula is “smaller” than the incorrect one! We will see this later when we cover stratified sampling and treatment effect estimation.

- To estimate the asymptotic variance of the two-step estimator, let

$$\hat{\mathbf{F}} = N^{-1}\sum_{i=1}^{N}\nabla_{\boldsymbol{\gamma}}\mathbf{s}_i(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\gamma}}) \xrightarrow{p} \mathbf{F}_o.$$

- Sometimes, replace $\nabla_{\boldsymbol{\gamma}}\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*)$ with $E[\nabla_{\boldsymbol{\gamma}}\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}^*)|\mathbf{x}_i]$ where $\mathbf{x}_i$ are conditioning variables (such as in NLS and conditional MLE).
- Let

$$\hat{\mathbf{g}}_i = \hat{\mathbf{s}}_i + \hat{\mathbf{F}}\hat{\mathbf{r}}_i$$

where estimates are evaluated at $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$.

- Notice that $\hat{\mathbf{g}}_i$ is an adjusted version of the score for the second-step problem. Then

$$\hat{\mathbf{D}} = N^{-1}\sum_{i=1}^{N}\hat{\mathbf{g}}_i\hat{\mathbf{g}}_i' \xrightarrow{p} \mathbf{D}_o.$$

- Also,

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N}\mathbf{H}_i(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\gamma}})$$

where $\mathbf{H}_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}_i(\boldsymbol{\theta};\boldsymbol{\gamma})$ is the $P \times P$ Hessian for the second-step problem. Can replace this with a Hessian conditional on $\mathbf{x}_i$.

- A valid estimator is then

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{D}}\hat{\mathbf{A}}^{-1}/N.$$

- If $\mathbf{F}_o = \mathbf{0}$, then

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N.$$

where $\hat{\mathbf{B}}$ is the usual outer product of the score (with respect to $\boldsymbol{\theta}$):

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'.$$

- If the scores from the two problems are uncorrelated,

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}^{-1}(\hat{\mathbf{B}} + \hat{\mathbf{F}}\hat{\mathbf{L}}\hat{\mathbf{F}}')\hat{\mathbf{A}}^{-1}$$

where

$$\hat{\mathbf{L}} = N^{-1}\sum_{i=1}^{N}\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i'.$$

- In the case of WNLS,

$$\nabla_{\boldsymbol{\theta}}\mathbf{s}_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \nabla_{\boldsymbol{\theta}}m(\mathbf{x}_i,\boldsymbol{\theta})'\nabla_{\boldsymbol{\theta}}h(\mathbf{x}_i,\boldsymbol{\gamma})u_i(\boldsymbol{\theta})/[h(\mathbf{x}_i,\boldsymbol{\gamma})^2]$$

so for any $\boldsymbol{\gamma}$,

$$\begin{aligned} E[\nabla_{\boldsymbol{\theta}}\mathbf{s}_i(\boldsymbol{\theta}_o,\boldsymbol{\gamma})] &= E\{\nabla_{\boldsymbol{\theta}}m(\mathbf{x}_i,\boldsymbol{\theta}_o)'\nabla_{\boldsymbol{\theta}}h(\mathbf{x}_i,\boldsymbol{\gamma})u_i/[h(\mathbf{x}_i,\boldsymbol{\gamma})^2]\} \\ &= \mathbf{0} \end{aligned}$$

because $E(u_i|\mathbf{x}_i) = 0$. So $\mathbf{F}_o = \mathbf{0}$ and we do not need to adjust for estimation of $\boldsymbol{\gamma}^*$.

- A robust variance matrix estimator that does not restrict $Var(y|\mathbf{x})$ is

$$\widehat{Avar}(\hat{\boldsymbol{\theta}}_{WNLS}) = \left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}}\hat{m}_i'\nabla_{\boldsymbol{\theta}}\hat{m}_i/\hat{h}_i\right)^{-1}\left(\sum_{i=1}^{N} \hat{u}_i^2\nabla_{\boldsymbol{\theta}}\hat{m}_i'\nabla_{\boldsymbol{\theta}}\hat{m}_i/\hat{h}_i^2\right)$$

$$\cdot\left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}}\hat{m}_i'\nabla_{\boldsymbol{\theta}}\hat{m}_i/\hat{h}_i\right)^{-1}$$

which looks like the robust NLS form except every appearance of $\hat{u}_i$ and $\nabla_{\boldsymbol{\theta}}\hat{m}_i$ is weighted by $1/\sqrt{\hat{h}_i}$.

7. BOOTSTRAPPING

- Sometimes for complicated estimation methods it is difficult to derive analytic formulas for quantities of interest. One might rather let a computer do the work. *Resampling methods* avoid applying the delta method (or other asymptotic tools) to obtain valid inference for various econometric procedures.
- In some cases, resampling actually improves on the standard $\sqrt{N}$-asymptotics. In others, it does not provide a formal improvement but often seems to work better in practice.

- Still need to assume the estimation problem is has some smoothness in the parameters. Easiest case is resampling under random sampling, which is our standard setting
- The *nonparametric bootstrap* is the most straightforward resampling scheme. The idea is to treat the observed data as a population, and resample from the sample. Let $\{\mathbf{w}_i : i = 1, \ldots, N\}$ be the realized sample, and suppose what we have an estimate, say $\hat{\gamma}$, based on this sample. Assuming $\hat{\gamma}$ is a smooth function of the data, how can we approximate a standard error for $\hat{\gamma}$?

- We repeatedly draw random samples from $\{\mathbf{w}_i : i = 1, \ldots, N\}$ of size N, which means sampling with replacement. In practice, one randomly draws N integers from $\{1, 2, \ldots, N\}$, with replacement, and these indices define a bootstrap sample of data.
- In effect, we treat $\{\mathbf{w}_i : i = 1, \ldots, N\}$ as the population and draw random samples from it.
- For a *bootstrap sample b*, denote the sample as $\{\mathbf{w}_1^{(b)}, \mathbf{w}_2^{(b)}, \ldots, \mathbf{w}_N^{(b)}\}$. Unless we draw each integer exactly once – thereby getting the original sample – a bootstrap sample will contain repeats of some observations and exclude others entirely.

- For bootstrap sample b, we use $\{\mathbf{w}_1^{(b)}, \mathbf{w}_2^{(b)}, \ldots, \mathbf{w}_N^{(b)}\}$ to obtain a set of estimates, say $\hat{\boldsymbol{\theta}}^{(b)}$. The estimate from the orginal sample is $\hat{\boldsymbol{\theta}}$.
- For a scalar estimate $\hat{\gamma} = g(\hat{\boldsymbol{\theta}})$ for a continuously differentiable function $g : \mathbb{R}^P \to \mathbb{R}$, we obtain its **bootstrap standard error** as

$$se_B(\hat{\gamma}) = \left[(B-1)^{-1} \sum_{b=1}^{B} (\hat{\gamma}^{(b)} - \bar{\hat{\gamma}})^2 \right]^{1/2}$$

where $\hat{\gamma}^{(b)} = g(\hat{\boldsymbol{\theta}}^{(b)})$ and $\bar{\hat{\gamma}} = B^{-1} \sum_{b=1}^{B} \hat{\gamma}^{(b)}$ is the average estimate across the bootstrap samples.

- We can use $se_B(\hat{\gamma})$ to construct asymptotic hypotheses tests and confidence intervals for γ_o based on the original estimate $\hat{\gamma}$.
- Especially for computing average partial effects – a topic that will arise repeatedly later – we often need to estimate a parameter that can be written as $\gamma_o = E[g(\mathbf{w}_i, \boldsymbol{\theta}_o)]$. A natural, consistent estimator is $\hat{\gamma} = N^{-1}\sum_{i=1}^{N} g(\mathbf{w}_i, \hat{\boldsymbol{\theta}})$. To estimate its asymptotic variance, we must account for the randomness in $\mathbf{w}_i$ as well as $\hat{\boldsymbol{\theta}}$. As before, we draw bootstrap samples, and, for bootstrap sample b, the estimate of γ_o is

$$\hat{\gamma}^{(b)} = N^{-1}\sum_{i=1}^{N} g(\mathbf{w}_i^{(b)}, \hat{\boldsymbol{\theta}}^{(b)}).$$

- Using the bootstrap standard error to construct test statistics cannot be shown to improve on the approximation provided by the usual asymptotic theory. As it turns out, in many cases the bootstrap *does* improve the approximation of the distribution of properly computed test statistics. In other words, the bootstrap can provide an **asymptotic refinement** compared with the usual asymptotic theory. But one must use some care in computing the bootstrap test statistics.

- In order to show that the bootstrap approximation of a distribution converges more quickly than the usual rates associated with first-order asymptotics, the notion of an **asymptotically pivotal statistic** is critical. An asymptotically pivotal statistic is one whose limiting distribution does not depend on unknown parameters.
- Asymptotic *t* statistics, Wald statistics, score statistics, and quasi-LR statistics are all asymptotically pivotal when they converge to the standard normal distribution (in the case of a *t* statistic) or the chi-square distribution in the case of the other statistics.

- One must be careful to ensure a statistic is asymptotically pivotal. For example, for a *t* statistic to be asymptotically pivotal in the context of nonlinear regression with heteroskedasticity, we must use a heteroskedasticity-robust statistic. The Wald and score statistics should use robust asymptotic variance estimators to generally deliver an asymptotic chi-square distribution. The quasi-LR statistic is guaranteed to be asymptotically pivotal only when the generalized information matrix equality holds.

- Consider testing $H_0 : \theta_o = c$ for some known value c. The t statistic, $t = (\hat{\theta} - c)/se(\hat{\theta})$ is asymptotically pivotal if $se(\hat{\theta})$ is appropriately chosen. In order to obtain a refinement using the bootstrap, we must obtain the empirical distribution of the statistic

$$t^{(b)} = (\hat{\theta}^{(b)} - \hat{\theta})/se(\hat{\theta}^{(b)})$$

where $\hat{\theta}$ is the estimate from the original sample, $\hat{\theta}^{(b)}$ is the estimate for bootstrap sample $b,$ and $se(\hat{\theta}^{(b)})$ is the standard error estimated from the same bootstrap sample. (So, for example, $se(\hat{\theta}^{(b)})$ could be a heteroskedasticity-robust standard error for nonlinear least squares.)

- Notice how the t statistic for each bootstrap replication is centered at the original estimate, $\hat{\theta}$, not the hypothesized value. As discussed by Horowitz (2001, *Handbook of Econometrics*, Volume 5), centering at the estimate is required to ensure asymptotic refinements of the testing procedure.
- Using the bootstrapped t statistics, we can obtain **bootstrap critical values**. We must decide on the nature of the alternative.

- For a one-sided alternative, say $H_0 : \theta_o > c$, we order the statistics $\{t^{(b)} : b = 1, 2, \ldots, B\}$, from smallest to largest, and we pick the value representing the desired quantile of the list of ordered values. For example, to obtain a 5% test against a greater than one-sided alternative , we choose the critical value as the 95^{th} percentile in the ordered list of $t^{(b)}$.

- For a two-sided alternative, we must choose between a **nonsymmetrical test** and a **symmetrical test**. For the former, a test with size α chooses critical values as the lower and upper $\alpha/2$ quantiles of the ordered bootstrap test statistics, and we reject H_0 if $t > cv_u$ or $t < cv_l$. For the latter, we first order the absolute values of the statsitics, $|t^{(b)}|$, and then choose the upper α quantile as the critical value for a test of size α. Naturally, we compare $|t|$ with the critical value. This approach to choosing critical values from bootstrapping is called the **percentile-*t* method**.

- We can use the percentile-t method to compute a **bootstrap *p*-value**. For example, against a greater than one-sided alternative, we simply find the fraction of bootstrap statistics $t^{(b)}$ that exceed t. A symmetric p-value for a two-sided alternative does the same for $|t^{(b)}|$ and $|t|$.

- Testing multiple hypotheses is similar. Suppose that for a Q –vector $\boldsymbol{\phi}_o$, we want to test $H_o : \boldsymbol{\phi}_o = \mathbf{r}$, where $\mathbf{r}$ is a vector of known constants. The Wald statistic computed using the original sample is $W = (\hat{\boldsymbol{\phi}} - \mathbf{r})'\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\phi}} - \mathbf{r})$. We compute a series of Wald statistics from bootstrap samples as:

$$W^{(b)} = (\hat{\boldsymbol{\phi}}^{(b)} - \hat{\boldsymbol{\phi}})'\left(\hat{\mathbf{V}}^{(b)}\right)^{-1}(\hat{\boldsymbol{\phi}}^{(b)} - \hat{\boldsymbol{\phi}}),\ b = 1,\ldots,B,$$

where we must take care so that the calculation of $\hat{\mathbf{V}}$ (and $\hat{\mathbf{V}}^{(b)}$) delivers an asymptotic chi-square statistic. The bootstrap p-value is the fraction of $W^{(b)}$ that exceed W.

- The nonparametric bootstrap applies directly to panel data settings with large N and small T. A draw $\mathbf{w}_i$ represents the data for all T time periods for unit i. That is, the resampling is of cross section units. Whenever we draw an index from $\{1, 2, \ldots, N\}$, we take all T time periods.
- Resampling cross section units with panel data is sometimes called the **panel bootstrap**.
- Resampling different time periods, as is done with pure time series applications and sometimes with panel data sets with small N and large T, is much harder and not appropriate for our setting.

APPLICATION: We use bootstrapping to obtain standard errors in the context of nonlinear least squares with cross section data, and compare the standard errors with those obtained from first-order asymptotics.

- Consider estimating a wage equation for hourly workers. We consider the standard linear model approach

$$\log(wage) = \alpha_0 + \alpha_1 female + \alpha_2 educ + \alpha_3 exper + \alpha_4 exper^2 + u$$

- Alternatively, we can directly estimate the (approximate) semi-elasticities on the conditional mean of *wage*:

$$E(wage|\mathbf{x}) = \exp(\beta_0 + \beta_1 female + \beta_2 educ + \beta_3 exper + \beta_4 exper^2)$$

- If we assume that u and $\mathbf{x}$ are independent in the log-level linear model, then the slopes in the two formulations are the same. If u has, say, heteroskedasticity, $E(wage|\mathbf{x})$ is not generally of the simple form above.

- We can turn that around, too. If we start with $E(wage|\mathbf{x})$ as the object of interest and specify it as an exponential form, there is no guarantee that a linear regression with $\log(wage)$ as the dependent variable consistently estimates the β_j.

```
. use wage1

. reg lwage female educ exper expersq, robust

Linear regression                                      Number of obs =     526
                                                       F(  4,   521) =   81.97
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3996
                                                       Root MSE      =  .41345

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3371868   .0361838    -9.32   0.000    -.4082709   -.2661026
        educ |   .0841361     .00769    10.94   0.000      .069029    .0992432
       exper |     .03891   .0046752     8.32   0.000     .0297253    .0480946
     expersq |   -.000686   .0001005    -6.83   0.000    -.0008834   -.0004887
       _cons |    .390483   .1085985     3.60   0.000     .1771383    .6038278
------------------------------------------------------------------------------

. * The estimates are pretty standard from log(wage) equations.
. * Now use NLS with an exponential mean function, and fully
. * robust standard errors.

. glm wage female educ exper expersq, fam(normal) link(log) robust

Generalized linear models                          No. of obs      =       526
Optimization     : ML                              Residual df     =       521
                                                   Scale parameter =   8.30647
Deviance         =  4327.670955                    (1/df) Deviance =   8.30647
Pearson          =  4327.670955                    (1/df) Pearson  =   8.30647

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  4.964372
Log pseudolikelihood = -1300.629849                BIC             =  1063.449

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3683686   .0538735    -6.84   0.000    -.4739588   -.2627784
        educ |   .1034196   .0120236     8.60   0.000     .0798537    .1269855
       exper |   .0494462   .0065125     7.59   0.000      .036682    .0622105
     expersq |  -.0008688   .0001415    -6.14   0.000    -.0011462   -.0005914
       _cons |    .137639   .1817583     0.76   0.449    -.2186007    .4938786
------------------------------------------------------------------------------
```

```
. * Now bootstrap the standard errors for the female and educ coefficients:

. do nls1

. capture program drop nls_boot

. 
. program nls_boot, rclass
  1. glm wage female educ exper expersq, fam(normal) link(log)
  2. return scalar bfemale = _b[female]
  3. return scalar beduc = _b[educ]
  4. 
. end

. 
. bootstrap r(bfemale) r(beduc), reps(1000) seed(123): nls_boot
(running nls_boot on estimation sample)

Bootstrap replications (1000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5 
..................................................    50
...
..................................................   950
..................................................  1000
```

```
Bootstrap results                               Number of obs     =        526
                                                Replications      =       1000

      command:  nls_boot
        _bs_1:  r(bfemale)
        _bs_2:  r(beduc)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |  -.3683686    .054011    -6.82   0.000    -.4742282    -.262509
       _bs_2 |   .1034196   .0122406     8.45   0.000     .0794285    .1274107
------------------------------------------------------------------------------

.
. program drop nls_boot

end of do-file
```

- In this application, we see that taking the log and using linear regression produces substantially smaller standard errors than using NLS on an exponential function. (This comparison only makes sense if we assume both methods are consistent for the parameters of interest.) But there are different ways to estimate an exponential mean that are more efficient than NLS. For example, later we will cover quasi-maximum likelihood in the linear exponential family.