ECON 7710

Econometrics I

Lecture notes 3.

Extremum estimation:

- Object of extremum estimation
  - Parameter of interest: $\theta \in \Theta \subset \mathbb{R}^p$
  - Work with convex compacts (usually)
  - Structural variable Y w. realizations y
  - Economic model $Y \sim F(\cdot, \theta)$
  - "True" DGP corresponds to $\theta = \theta_0 \in int\,(\Theta)$
  - Function $Q(\theta) = E_{\theta_0}[g\,(Y, \theta)] = \int g\,(y, \theta)\,F(dy, \theta_0)$
  - Note: integrate against true distribution
  - Extremum estimation:
  $$\theta_0 = \text{argmax}_{\theta \in \Theta} Q(\theta)$$
  - So far we don't know where $g(\cdot)$ is coming from
- Example (OLS)
  - Observable structural variables W, X
  - Dgp
  $$W = X'\theta + \varepsilon$$
  - $E[\varepsilon] = 0$, $E[\varepsilon^2] = \sigma^2$
  - $g(w, x; \theta) = \varepsilon^2 = (w - x'\theta)^2$
  - $Q(\theta) = E\left[(w - x'\theta)^2\right]$
  - Computing true expectation is not feasible

- Why? The distribution $F(\cdot)$ is not known (because θ_0 is not known!)
- So, somehow, need to approximate expectaion $E_{\theta_0}[]$ without knowing the true parameter
- Have sample $y_1, \ldots, y_T$ (i.i.d.)
- Note: $\frac{1}{T}\sum\limits_{t=1}^{T} y_t \xrightarrow{p} E_{\theta_0}[Y]$
- This seems to give a solution!

- Analogy principle
  - Use sample analog to approximate expectation:
$$\widehat{Q}(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(y_t;\theta) \equiv E_T\left[g(Y;\theta)\right]$$
  - Define sample analog
$$\hat{\theta} = \text{argmax}_{\theta\in\Theta}\widehat{Q}(\theta)$$
  - How close is $\hat{\theta}$ to θ_0?
  - Note that have two pieces: function of θ and approximation of expectation by sample sum
  - Need convergence concept to see approach of $\hat{\theta}$ to θ_0

- **Definition:** Let $\{Q_T(\theta)\}_{T=1}^{\infty}$ -non-negative sequence of random functions. Then if
  - (i) $\Pr\left(\lim\limits_{T\to\infty}\sup\limits_{\theta\in\Theta} Q_T(\theta) = 0\right) = 1$ then $Q_T(\theta)$ converges to 0 a.s. uniformly in θ
  - (ii) For any $\varepsilon > 0$ $\lim\limits_{T\to\infty}\Pr\left(\sup\limits_{\theta\in\Theta} Q_T(\theta) < \varepsilon\right) = 1$ then $Q_T(\theta)$ converges to 0 in probability uniformly in θ

- **Theorem:** Assume that
  - (a) Θ is compact
  - (b) $\widehat{Q}_T(\theta)$ is continuous in Θ
  - (c) $\widehat{Q}_T(\theta)$ converges in probability to $Q(\theta)$ uniformly in Θ

- (d) $Q(\cdot)$ attains a unique global maximum at θ_0 (identification)

Then if $\hat{\theta} = \text{argmax}_{\theta \in \Theta} \widehat{Q}_T(\theta)$ then $\hat{\theta} \xrightarrow{p} \theta_0$

- Example (NLLS)
  - $W = m(X, \theta) + \varepsilon$, function $m(\cdot)$ is known
  - $m(\cdot)$ is twice differentiable in θ
  - Conditions $E[\varepsilon] = 0$, and $E[\varepsilon^2] < \infty$
  - Objective $Q(\theta) = E\left[(W - m(X, \theta))^2\right]$
  - Identification (local): necessary and sufficient conditions for minimum are satisfied
  - Necessary condition: $\frac{\partial}{\partial \theta} Q(\theta_0) = 0$
  - $Q(\theta) = \int \int (w - m(x, \theta))^2 f(w, x; \theta_0)\, dw\, dx$
  - FOC:
  $$\frac{\partial}{\partial \theta} Q(\theta_0) = -2 \int \int (w - m(x, \theta)) \frac{\partial m}{\partial \theta} f(w, x; \theta_0)\, dw\, dx$$
  - SOC:
  $$l \frac{\partial^2}{\partial \theta^2} Q(\theta_0) = -2 \int \int \left[(w - m(x, \theta)) \frac{\partial^2 m}{\partial \theta^2} \right. \tag{1}$$
  $$\left. - \left(\frac{\partial m}{\partial \theta}\right)^2 \right] f(w, x; \theta_0)\, dw\, dx \tag{2}$$
  $$\tag{3}$$
  - Identification condition:
    - Equation $E\left[(w - m(x, \theta)) \frac{\partial m}{\partial \theta}\right] = 0$ has a unique solution
    - OR $E\left[(w - m(x, \theta)) \frac{\partial^2 m}{\partial \theta^2} - \left(\frac{\partial m}{\partial \theta}\right)^2\right] < 0$

    at point θ_0
- Sample analog $\widehat{Q}(\theta) = \frac{1}{T} \sum_{t=1}^{T} (w_t - m(x_t, \theta))^2$
- Verify conditions of theorem?
- (a) and (b) are satisfied automatically

- (d) is satisfied if SOC holds
- (c) can be tricky
- Rough idea to prove uniform convergence is to slice the parameter space and show convergence in slices
- Very tedious. In the rest of the course we just assume uniformity
- In real problems have pre-packaged results (HE)
- From out Theorem conclude that minimizer of sample NLLS will converge to population NLLS
- Turns out that can also provide asymptotic results
- Asymptotic distribution
  - Mean-value expansion: main work tool!
  - From FOC
$$\frac{\partial \widehat{Q}(\hat{\theta})}{\partial \theta} = 0$$
  - Mean-value expansion at θ_0
$$\frac{\partial \widehat{Q}(\theta_0)}{\partial \theta} + \frac{\partial^2 \widehat{Q}(\theta^*)}{\partial \theta^2}(\hat{\theta} - \theta_0)$$
  - Know that $\frac{\partial Q(\theta_0)}{\partial \theta} = 0$
  - $\sqrt{T}\frac{\partial \widehat{Q}(\theta_0)}{\partial \theta} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial g(y_t;\theta_0)}{\partial \theta}$
  - CLT
$$\frac{1}{\sqrt{T}} \frac{\partial g(y_t;\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \Sigma)$$
  - LLN and coninuous mapping
$$\frac{\partial^2 \widehat{Q}(\theta^*)}{\partial \theta^2} \xrightarrow{p} \frac{\partial^2 Q(\theta_0)}{\partial \theta^2}$$
  - $\hat{\theta} - \theta_0$ - asymptotically normal

- **Theorem:** Assume (a)-(d) in the previous theorem and
  - (e) $\frac{\partial^2 \widehat{Q}}{\partial\theta\partial\theta'}$ exists in the neighborhood of θ_0
  - (f) $\frac{\partial^2 \widehat{Q}(\theta_T)}{\partial\theta\partial\theta'} \xrightarrow{p} A(\theta_0)$ for any $\theta_T \xrightarrow{p} \theta_0$
  - (g) $\sqrt{T}\frac{\partial \widehat{Q}(\theta_0)}{\partial\theta} \xrightarrow{d} N(0, B(\theta_0))$

  Then if $\hat{\theta}$ solves FOC, then

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A(\theta_0)^{-1\prime} B(\theta_0) A(\theta_0)^{-1})$$

Maximum likelihood:

- MLE assumptions
  - $Y \sim F(\cdot, \theta_0)$
  - y_t are i.i.d.
  - $F(\cdot, \theta)$ is parametrized by $\theta \in \Theta \subset \mathbb{R}^p$
  - Distribution $F(\cdot, \theta)$ is dictated by our economic model (examples below)
- MLE objective
  - "Single observation" likelihood function: density of Y
  - Log-likelihood of single observation: $l(Y, \theta) = \log\, f(Y, \theta)$
  - Population objective function

$$L(\theta) = E_{\theta_0}\left[\log\, f(Y, \theta)\right] = \int \log\, f(y, \theta)\, f(y, \theta_0)\, dy$$

  - Search to maximize this objective
- MLE as "distance" minimization
  - Consider objective $-L(\theta) = E_{\theta_0}\left[-\log\, f(Y, \theta)\right]$
  - This needs to be minimized
  - Now consider **constant**: $E_{\theta_0}\left[\log\, f(Y, \theta_0)\right]$ (fixed function integrated against fixed distribution)

- Define objective:

$$KL(\theta) = E_{\theta_0}\left[\log\, f(Y,\theta_0)\right] - L(\theta) = E_{\theta_0}\left[\log\, \frac{f(Y,\theta_0)}{f(Y,\theta)}\right]$$

- This is called Lullback-Leibler information (KLIC)
- This is "distance" between true and estimated distributions that we minimize
- Not true distance because it is asymmetric

- Sample analogs
  - KL cannot be used for estimation directly because θ_0 (and $f(Y,\theta_0)$) are unknown
  - As before, use sample analog $\frac{1}{T}\sum_{t=1}^{T}$ to approximate expectation
  - Construct sample log-likelihood function

$$\widehat{L}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\log\, f(y_t,\theta)$$

  - Find sample analog estimate

- Example: Linear regression
  - $W = X'\beta + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$
  - $\theta = \left(\beta,\sigma^2\right)$
  - Conditional density of dependent variable

$$f\left(W \mid x;\beta,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(w-x'\theta)}{2\sigma^2}\right)$$

  - Log-likelihood in the sample

$$\widehat{L}(\beta,\sigma^2) = -\frac{T}{2}\log\left(2\pi\sigma^2\right) - \frac{(w-x'\theta)}{2\sigma^2}$$

- Example: Discrete choice
  - Unobserved utility: $W = X'\beta + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$

- Make choice if utility is positive

$$D = \mathbf{1}\{W > 0\}$$

- Observe D and X: choices and covariates
- Now estimate only one parameter β (soon we will see why!)
- Conditional distribution of outcome is discrete:

$$\Pr\left(D = 1 \,|\, X = x, \beta\right) = P\left(x'\beta + \varepsilon > 0\right) = 1 - \Phi(-x'\beta) = \Phi(x'\beta)$$

- Make choice if utility is positive

$$D = \mathbf{1}\{W > 0\}$$

- Observe D and X: choices and covariates
- Now estimate only one parameter β (soon we will see why!)
- Log-likelihood in the sample

$$\widehat{L}(\beta) = \sum_{t=1}^{T} d_t \log\, \Phi(x'\beta) + (1 - d_t) \log\left(1 - \Phi(x'\beta)\right)$$

- **Definition:** The likelihood function of a random variable Y with density $f(\cdot, \theta)$ is a function of parameter θ: $l(\theta; y) = f(y, \theta)$
  Log-likelihood function: $L(\theta; y) = \log\, l(\theta; y)$
  Conditional likelihood: $l(\theta; w|z) = f(w, \theta|z)$
  - For discrete distribution, use the pmf instead of pdf
  - Likelihood function is function of parameters, provided the sample
  - Interpretation: Maximize the probability of observing a given sample of data
- **MLE assumptions**
  - $Y \sim f(\cdot, \theta_0)$, i.i.d., $\theta \in \Theta$ - convex compact set
  - $E\left\{\sup_{\theta\in\Theta} |\log\, f(Y, \theta)|\right\} < \infty$ (Note: expectation is taken wrt to $f(\cdot, \theta_0)$)

- $\log f(y_t, \theta)$ is continuous in θ

- **Lemma:** $E\left[\log f(Y,\theta)\right] \leq E\left[\log f(Y,\theta_0)\right]$

- *Proof:* For concave $h(\cdot)$: $E[h(Y)] \leq h\left(E[Y]\right)$. As a result:

$$E\left[\log \frac{f(Y,\theta)}{f(Y,\theta_0)}\right] \leq \log E\left[\frac{f(Y,\theta)}{f(Y,\theta_0)}\right].$$

  Note $E\left[\frac{f(Y,\theta)}{f(Y,\theta_0)}\right] = \int \frac{f(y,\theta)}{f(y,\theta_0)} f(y,\theta_0)\, dy = 1$ Thus $E\left[\log \frac{f(Y,\theta)}{f(Y,\theta_0)}\right] \leq 0$. This proves the lemma

- Takeaway points
  - Population log-likelihood takes the highest value at true parameter value
  - This justifies why we focus on maximum likelihood
  - KLIC is always non-negative
  - We would also prefer that KLIC=0 if and only if $\theta = \theta_0$
  - This is actually required for identification!

- Dependence on support
  - Condition that $E\left\{\sup_{\theta\in\Theta} |\log f(Y,\theta)|\right\} < \infty$ is very important
  - This condition is violated when support of $f(\cdot)$ depends on θ
  - This could be bad: violated in case of uniform distribution
  - This case is called "superconsistent" MLE case

- Example
  - Exponential distribution

$$f(y,\theta) = \begin{cases} 0, & \text{if } y < \theta, \\ \exp\left(-(y-\theta)\right),; & \text{if } y \geq \theta. \end{cases}$$

  - Leads to log-likelihood

$$\log f(y,\theta) = \begin{cases} -\infty, & \text{if } y < \theta, \\ -(y-\theta),; & \text{if } y \geq \theta. \end{cases}$$

- $\sup_{\theta\in\Theta} |\log\ f(y,\theta)| \to +\infty$
- Our assumption is violated
- Possible solution: pick the parameter space $\Theta = (-\infty, \theta_0]$
- Then $E\left\{\sup_{\theta\leq\theta_0} |\log\ f(Y,\theta)|\right\} < \infty$
- Not feasible: don't know θ_0!

- **Definition:**
  - (i) Population likelihood function $L(\theta) = E\left[\log\ f(Y,\theta)\right]$
  - (ii) Sample likelihood function (for i.i.d. data) $\widehat{L}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\log\ f(y_t,\theta)$
  - (iii) Maximum likelihood estimator $\hat{\theta}_{MLE} = \text{argmax}_{\theta\in\Theta}\widehat{L}(\theta)$

- Example: Discrete choice model
  - $W = X'\beta + \varepsilon$, $\varepsilon \sim N(0,1)$
  - $D = \mathbf{1}\{W > 0\}$
  - $P\left(D = d|x\right) = \Phi\left(x'\beta\right)^d\left(1 - \Phi\left(x'\beta\right)\right)^{1-d}$
  - Conditional log-likelihood (one element)

$$\log\ P\left(D = d|x\right) = d\log\ \Phi\left(x'\beta\right) + (1-d)\log\ \left(1 - \Phi\left(x'\beta\right)\right)$$

  - For full likelihood: need also density of X: $f_X(\cdot)$ (assume known)
  - Full likelihood is separable in X and D distributions
  - Population conditional log-likelihood

$$L(\theta|x) = E\left[\log\ P\left(D = d|x\right)|x\right]$$

  - Thus

$$L(\theta|x) = \Phi\left(x'\beta\right)\log\ \Phi\left(x'\beta\right) + \left(1 - \Phi\left(x'\beta\right)\right)\log\ \left(1 - \Phi\left(x'\beta\right)\right)$$

  - Full likelihood

$$L(\theta) = E\left[\Phi\left(X'\beta\right)\log\ \Phi\left(X'\beta\right) + \left(1 - \Phi\left(X'\beta\right)\right)\log\ \left(1 - \Phi\left(X'\beta\right)\right)\right]$$

- Sample log-likelihood function

$$\widehat{L}(\theta) = \frac{1}{T}\sum_{t=1}^{T} d_t \log\, \Phi\left(x_t'\beta\right) + (1-d_t)\log\,\left(1 - \Phi\left(x_t'\beta\right)\right)$$

- Maximization requires that FOC is satisfied

$$\frac{\partial \widehat{L}(\theta)}{\partial \theta} = \frac{1}{T}\sum_{t=1}^{T} \frac{d_t - \Phi(x_t'\beta)}{\Phi(x_t'\beta)\left(1-\Phi(x_t'\beta)\right)}\phi(x_t'\beta)x_t = 0.$$

- Log-likelihood function is globally concave, thus the MLE estimator is a unique maximizer

- **Definition:** Suppose that $Y \sim f(\cdot, \theta_0)$ are i.i.d. and $\theta \in \Theta$. Then parameter θ_0 is not identified is there exists θ^* such that $\theta^* \neq \theta_0$ and $L(\theta_0) = E\left[\log\, f(Y,\theta_0)\right] = L(\theta^*) = E\left[\log\, f(Y,\theta^*)\right]$

- Identification
  - Note that we don't need $f(y,\theta_0) \equiv f(y,\theta^*)$ for lack of identification
  - Provided that our information is coming from distribution, cannot distinguish parameters and functions that lead to the same result on different distributions
  - Non-identification in the sense of previous definition is sometimes called global non-identification
  - Natural definition of identification
  - Parameter θ_0 is identified in Θ if for all $\theta \in \Theta$ and $\theta \neq \theta_0$

$$\Pr\left\{\log f(Y,\theta_0) \neq \log\, f(Y,\theta^*)\right\} > 0$$

- Example: discrete choice model
  - $W = X'\beta + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$
  - $D = \mathbf{1}\{W > 0\}$
  - New parameter $\theta = (\beta, \sigma^2)$

- This new parameter is **NOT** identified
- $P(D=d|x) = \Phi(\frac{x'\beta}{\sigma})^d(1-\Phi(\frac{x'\beta}{\sigma}))^{1-d}$
- Pick new parameter $\beta^* = \alpha\beta_0$ and $\sigma^* = \alpha\sigma_0$
- Then

$$\Phi\left(\frac{x'\beta^*}{\sigma^*}\right)^d\left(1-\Phi\left(\frac{x'\beta^*}{\sigma^*}\right)\right)^{1-d}$$
$$\equiv$$
$$\Phi\left(\frac{x'\beta_0}{\sigma_0}\right)^d\left(1-\Phi\left(\frac{x'\beta_0}{\sigma_0}\right)\right)^{1-d}$$

- $W = X'\beta + \varepsilon$, $\varepsilon \sim N(0,1)$
- $D = \mathbf{1}\{W > 0\}$
- New parameter $\theta = (\beta)$
- This new parameter is globally identified
- $P\left(D=d|x\right) = \Phi\left(x'\beta\right)^d\left(1-\Phi\left(x'\beta\right)\right)^{1-d}$
- Given that log-likelihood is globally concave, it has a unique global maximum
- As a result if $|\beta^* - \beta_0| > \varepsilon$ then

$$\log\, P(D=d|x;\beta_0) - \log\, P(D=d|x;\beta^*) > 0$$

- Equality possible only when $\beta^* \equiv \beta_0$

- **Theorem:** Under MLE Assumptions and provided that $\theta_0 \in \text{int}(\Theta)$ is identified it follows that $\theta \neq \theta_0$ implies

$$L(\theta) = E\left[\log\, f(Y,\theta)\right] < E\left[\log\, f(Y,\theta_0)\right] = L(\theta_0)$$

- **Assumption:** $f(y,\theta)$ is twice continuously differentiable in Θ and the support of $f(\cdot)$ does not depend on θ. We also assume that Fubbini theorem can be applied and the differentiation can be taken inside the integral

$$\frac{\partial}{\partial\theta}\int f(y,\theta)\, f(y,\theta_0)\, dy = \int \frac{\partial f(y,\theta)}{\partial\theta}\, f(y,\theta_0)\, dy$$
$$\frac{\partial^2}{\partial\theta^2}\int f(y,\theta)\, f(y,\theta_0)\, dy = \int \frac{\partial^2 f(y,\theta)}{\partial\theta^2}\, f(y,\theta_0)\, dy$$

- For twice continuously differentiable objectives finding the maximum of $L(\theta)$ can be represented by the solution of FOC
- As a result, find the roots of FOC
- Also from the previous assumption this will be equivalent to finding roots of

$$E\left[\frac{\partial \log f(Y,\theta)}{\partial \theta}\right] = 0$$

  in Θ

- **Definition:** The score function

$$s(\theta, y) = \frac{\partial \log f(y,\theta)}{\partial \theta}$$

  is the gradient of the log-likelihood

- **Lemma:** Under Assumptions 1 and 2

$$E\left[s(\theta, y)\right] = 0.$$

- *Proof:* Note that $\int f(y,\theta)dy = 1$. Thus

$$\int \frac{\partial}{\partial \theta} f(y,\theta)dy = 0 = \int \frac{\frac{\partial f(y,\theta)}{\partial \theta}}{f(y,\theta)} f(y,\theta)dy = E[s(\theta, y)].$$

- **Definition:** Information of the model

$$I_\theta = \mathrm{Var}\left(s(\theta, y)\right)$$

  is the variance of the score

- I_θ is also called the information matrix. We will deal with cases $\|I_\theta\| > 0$, where $\|\cdot\|$ is the defined as the smallest eigenvalue.
- Example: Estimating the mean
  - Model

$$w_t = \alpha\beta + \alpha\varepsilon_t, \quad \varepsilon \sim N(0, \sigma^2)$$

- Can we identify α, β and σ^2? No!
- Log-likelihood

$$l(\theta) = -\frac{1}{2}\log(2\pi\sigma^{2\alpha^2}) - \frac{(w-\alpha\beta)^2}{2\alpha^2\sigma^2}$$

- We note that parameters α and σ^2 deliver the same log-likelihood value as $k\alpha$ and σ^2/k^2
- Compute the score
- $\frac{\partial l(\theta)}{\partial \alpha} = \frac{\varepsilon^2-\sigma^2}{\alpha\sigma^2} + \frac{\beta\varepsilon}{\alpha\sigma^2}$
- $\frac{\partial l(\theta)}{\partial \beta} = \frac{\varepsilon}{\alpha\sigma^2}$
- $\frac{\partial l(\theta)}{\partial \sigma} = \frac{\varepsilon^2-\sigma^2}{\sigma^3}$
- Score

$$s(\theta) = \begin{pmatrix} \beta & \frac{1}{\alpha} \\ 1 & 0 \\ 0 & \frac{1}{\sigma} \end{pmatrix} \begin{pmatrix} \frac{\varepsilon}{\alpha\sigma^2} \\ \frac{\varepsilon^2-\sigma^2}{\sigma^2} \end{pmatrix}$$

- Note that $E[\varepsilon] = 0$ and $E[\varepsilon^2 - \sigma^2] = 0$
- Thus $E[s(\theta, y)] = 0$ (our lemma is valid!)
- Information

$$I_\theta = \mathrm{Var}(s(\theta, y)) = \begin{pmatrix} \beta & \frac{1}{\alpha} \\ 1 & 0 \\ 0 & \frac{1}{\sigma} \end{pmatrix} \begin{pmatrix} \frac{1}{\alpha^2\sigma^2} & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \beta & 1 & 0 \\ \frac{1}{\alpha} & 0 & \frac{1}{\sigma} \end{pmatrix}$$

- I_θ is a 3 by 3 matrix and the expression above is its eigenvalue decomposition
- Only 2 eigenvectors and 2 eigenvalues
- The third eigenvalue is equal to zero!
- Models that are not identified have a singular information matrix

• Identification and information

  - Studying the rank of information matrix is extremely important in applied research!

- If information of your model is singular - alarming fact: (1) Think about your data; (2) Think about your model
- The relationship between identification and singularity of information of the model is not one-to-one
- If the model is not identified, information is singular
- If information is singular, it does not necessarily means that the model is not identified
- If the model has singular information, can conclude that the model cannot be estimated at $\sqrt{T}$-rate
- Recent studies show that many familiar models have singular information (treatment effects with unbounded support for conditional treatment probability)

- **Lemma:** If $Y \sim F(\cdot, \theta_0)$, regularity conditions are satisfied and the information matrix is non-singular, then

$$E\left[\frac{\partial^2 \log\, f(Y,\theta)}{\partial\theta\partial\theta'}\right] = -I_\theta$$

- *Proof:* We already know that $E\left[s(\theta, Y)\right] = 0$. We also know that $\frac{\partial^2 \log\, f(y,\theta)}{\partial\theta\partial\theta'} = \frac{\partial s(\theta,y)}{\partial\theta'}$. Therefore

$$\begin{aligned}&\tfrac{\partial}{\partial\theta'}\int s(\theta,y)f(y,\theta)\,dy = \int \tfrac{\partial s(\theta,y)}{\partial\theta'}f(y,\theta)\,dy\\ &+\int s(\theta,y)s(\theta,y)'dy = 0\end{aligned}$$

This means that

$$E\left[\frac{\partial^2 \log\, f(Y,\theta)}{\partial\theta\partial\theta'}\right] = -E\left[s(\theta,Y)s(\theta,Y)'\right] = -I_\theta.$$

- Remark
  - This result can be useful for maximum search
  - When sample likelihood is very sensitive to parameters, information provides a more robust estimate for the Hessian
  - Need that when search for maximum, e.g. using Newton-Raphson algorithm
  - Maximum search is based on solving FOC

$$\frac{\partial\widehat{L}(\theta^*)}{\partial\theta} = 0$$

- Mean-value expansion
$$\frac{\partial \widehat{L}(\tilde{\theta})}{\partial \theta} + \frac{\partial^2 \widehat{L}(\bar{\theta})}{\partial \theta \partial \theta'}(\tilde{\theta} - \theta^*) = 0$$
- Then use this formula to iterate: from starting point $\theta^{(0)}$ to point $\theta^{(k)}$:
$$\frac{\partial \widehat{L}(\theta_{k-1})}{\partial \theta} + \frac{\partial^2 \widehat{L}(\theta_{k-1})}{\partial \theta \partial \theta'}(\theta_k - \theta_{k-1}) = 0$$
- Move from θ_{k-1} to θ_k and iterate till the points become close
- If likelihood is very sensitive to parameters, computing the second derivative can be tricky
- Do not need evaluation of the Hessian, as we can use the information matrix as a good estimate for the Hessian!

- **Theorem:** Consider an economic model with $Y \sim f(\cdot, \theta_0)$. Suppose that $\hat{\theta}_T$ is an unbiased estimator for θ_0 ($E\left[\hat{\theta}_T\right] = \theta_0$). Under our regularity conditions we have:
$$\text{Var}(\sqrt{T}\left(\hat{\theta}_T - \theta_0\right)) \geq I_\theta^{-1}.$$
Here $\geq$ denotes that matrix $\text{Var}(\sqrt{T}\left(\hat{\theta}_T - \theta_0\right)) - I_\theta^{-1}$ is positive semidefinite (for diagonal matrices this is an element-by-element inequality)

- Cramer-Rao lower bound
  - This inequality establishes the Cramer-Rao lower bound
  - It shows the fundamental role of information for regular models: it establishes the lowest bound for the variance of regular estimator
  - Regularity here is important
  - In case of superefficient estimators can definitely beat this lower bound (by a lot!)
  - As we will see, this also implies very nice properties of the maximum likelihood estimator

- *Proof:* We start off from the definition of unbiasedness. Provided that the estimator is unbiased, for any DGP parametrized by θ
$$E_\theta\left[\hat{\theta}_T\right] = \theta$$

Note that $\hat{\theta}_T$ is the function of the sample $y_1, \ldots, y_T$.

In other words if the sample is i.i.d.

$$\int \hat{\theta}_T f(y_1, \theta) \ldots f(y_T, \theta) dy_1 \ldots dy_T = \theta$$

Differentiate this w.r.t. θ:

$$\sum_{t=1}^{T} \int \hat{\theta}_T f(y_1, \theta) \ldots \frac{\partial f(y_t, \theta)}{\partial \theta} \ldots f(y_T, \theta) dy_1 \ldots dy_T = I,$$

where I is the identity matrix.

Next, note that

$$\sum_{t=1}^{T} \int \hat{\theta}_T f(y_1, \theta) \ldots \frac{\partial f(y_t, \theta)}{\partial \theta} \ldots f(y_T, \theta) dy_1 \ldots dy_T = TE \left[\hat{\theta}_T s(\theta, y_t) \right]$$

Given that

$$\sum_{t=1}^{T} \int \hat{\theta}_T f(y_1, \theta) \ldots \frac{\partial f(y_t, \theta)}{\partial \theta} \ldots f(y_T, \theta) dy_1 \ldots dy_T = TE \left[\hat{\theta}_T s(\theta, y_t) \right]$$

we find that

$$\operatorname{cov}\left(\hat{\theta}_T, s(\theta, y_t) \right) = \frac{1}{T} I$$

Consider

$$Z = \begin{pmatrix} \sqrt{T} \left(\hat{\theta}_T - \theta \right) \\ \sum_{t=1}^{T} s(\theta, y_t) \end{pmatrix}$$

Note that $\operatorname{Var}(Z)$ is positive semidefinite (as covariance matrix)

$$\operatorname{Var}(Z) = \begin{pmatrix} \operatorname{Var}(\sqrt{T} \left(\hat{\theta}_T - \theta \right)) & T\sqrt{T} \operatorname{cov}(\hat{\theta}_T, s(\theta, y_t)) \\ T\sqrt{T} \operatorname{cov}(\hat{\theta}_T, s(\theta, y_t)) & TI_\theta \end{pmatrix}$$

Pick

$$c = \begin{pmatrix} -I \\ \frac{1}{\sqrt{T}} I_\theta^{-1} \end{pmatrix}$$

Given that $\operatorname{Var}(Z)$ is positive semidefinite

$$c' \operatorname{Var}(Z) c \geq 0$$

Then

$$
\begin{aligned}
& c'\mathrm{Var}(Z)c \\
&= \begin{pmatrix} -I & \frac{1}{\sqrt{T}} I_\theta^{-1} \end{pmatrix} \begin{pmatrix} \mathrm{Var}(\sqrt{T}\left(\hat{\theta}_T - \theta\right)) & \sqrt{T} I \\ \sqrt{T} I & T I_\theta^{-1} \end{pmatrix} \begin{pmatrix} -I \\ \frac{1}{\sqrt{T}} I_\theta^{-1} \end{pmatrix} \\
&= \mathrm{Var}(\sqrt{T}\left(\hat{\theta}_T - \theta\right)) - I_\theta^{-1} \geq 0
\end{aligned}
$$

This delivers the result of the theorem

- **Definition:** A consistent estimator is called (asymptotically) efficient if $\lim_{T\to\infty} \mathrm{Var}(\hat{\theta}_T) = I_\theta^{-1}$
- **Theorem:** Under our regularity conditions, the maximum likelihood estimator is asymptotically efficient.